Morality and Blame

Samuel Mason

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Philosophy, Religion and History of Science

May 2023

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I am very grateful to my supervisors, Gerald Lang and Pekka Väyrynen, for all of their help, advice, and encouragement. I would also like to thank the members of the Center for Aesthetic, Moral and Political Philosophy at the University of Leeds, and in particular Jess Isserow and Rich Rowland, for helpful discussions of some of the ideas in this thesis.

I am also grateful to the philosophy postgraduate community at the University of Leeds for a stimulating and friendly environment during my PhD studies, in particular: Ludovica Adamo, Kieran Bateman, Miriam Bowen, Radu Bumbăcea, Pei-Lung Cheng, Matt Clark, Aleksander Domoslawski, Jakob Donskov, Simon Graf, Maddy Page, Katie Prosser, Rodrigo Valencia-Pacheco, Jyl Schacher, and Alexios Stamatiadis-Bréhier.

Finally, I am grateful to my family, Lynn, Andy, Matt, and Evie, for all of their support throughout my PhD studies.

Abstract

This thesis argues that moral wrongness, permissibility, and requirement are conceptually and metaphysically analysable in terms of moral blameworthiness. As formulated in terms of moral wrongness, the analysis I defend holds:

> *Moral Wrongness as Moral Blameworthiness (MB)*: It is morally wrong for an agent to φ iff (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, where 'φ' stands for an object of deontic moral assessment.

Chapters 1-3 explain MB, situate it in relation to the sentimentalist tradition in meta-normative theory, and present my main positive arguments for MB. Chapters 4-6 respond to objections to MB. Finally, Chapter 7 examines the relation between moral wrongness and normative reasons in light of MB. I argue that MB supports the claim that we always have strong normative reasons not to act morally wrongly, but not any stronger claims concerning the connection between moral wrongness and normative reasons.

Contents

Chapter 1

Introduction

This thesis defends conceptual and metaphysical analyses of moral wrongness, permissibility, and requirement. The analyses belong to the sentimentalist tradition in meta-normative theory. As formulated in terms of moral wrongness, the analysis I defend holds:

> *Moral Wrongness as Moral Blameworthiness (MB)*: It is morally wrong for an agent to φ iff (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, where 'φ' stands for an object of deontic moral assessment.

As understood here, an agent is morally blameworthy for φ-ing just in case they would be a fitting target of moral blaming emotions for φ-ing. 'Moral blaming' emotions consist in guilt, resentment, and indignation.[1]

Although the statement of MB given above is formulated in terms of moral wrongness, MB can also be formulated in terms of moral requirement and moral permissibility. As formulated in terms of moral requirement, MB holds that an agent is morally required to φ iff (Def) not φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them. And as stated in terms of moral permissibility, MB holds that it is morally permissible for an agent to φ iff (Def) φ-ing does not violate standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them.

[1] I defend these claims about moral blame and moral blameworthiness, and explain what commitments they involve, in Chapter 2.

Similar analyses to MB have been defended by Mill (1861: Chapter 5), Gibbard (1990: 36-48, 1992, 2006), Skorupski (1993, 2010: 290-320), Darwall (2006a: 91-118, 2013a: 3-19), and Kauppinen (2017). Some of these philosophers defend MB (or something close to it) only as a conceptual analysis; others defend it only as a metaphysical analysis.[2] My work builds on the work of these philosophers by providing new arguments for MB, new replies to objections to MB, and a new account of the upshots of MB for the relation between moral wrongness and normative reasons.

This Introduction is split into three sections. Section 1.1 explains how I understand conceptual and metaphysical analysis. Section 1.2 locates my project among some of the extant work on sentimentalism in meta-normative theory. Finally, Section 1.3 provides chapter summaries.

1.1 Methodology

I begin (in 1.1.1) by explaining how I understand conceptual analysis. Two common objections to the method of conceptual analysis arise from the claim that there are no analytic truths, or at least no philosophically interesting analytic truths, and the 'Open Question Argument'. Although the conceptual analyses I defend aim to be broadly in line with the deontic moral concepts we in fact use, they are not purely descriptive: they are partly prescriptive or ameliorative proposals, justified partly by the claim that the account of deontic moral concepts provided by them explains how deontic moral concepts serve valuable social functions.[3] Within this broad framework, I give two more detailed models of conceptual analysis that occupy this

[2] Darwall is perhaps best read not as arguing that moral wrongness is conceptually reducible to moral blameworthiness, but rather that MORALLY WRONG, MORALLY BLAMEWORTHY, and various other 'second-personal' concepts form an interdependent and mutually illuminating set (cf. 2006a: 11-15).

[3] Insofar as my defence of MB has an ameliorative aspect, it involves what has been called 'conceptual engineering' (for a collection of perspectives on conceptual engineering, see Burgess et al. (2020)). I say more about this aspect of my defence of MB below.

territory between description and prescription. The first is developed from David Lewis's (1989) work, and the second from Allan Gibbard's (1990).[4] The first model purports to yield conceptual analyses that state analytic truths, whereas the second model purports to yield conceptual analyses that do not state analytic truths. My defence of MB is consistent with either of these models, and hence is consistent with either the acceptance of the claim that there are philosophically significant analytic truths or its rejection. Finally, I argue that both models of conceptual analysis have ample resources for responding successfully to the Open Question Argument.

I then (in 1.1.2) explain how I understand metaphysical analysis, show what account of the structure of the deontic moral domain results from MB as a metaphysical analysis, and contrast this account with some alternative views.

1.1.1 Conceptual Analysis

Analytic truths, as understood here, are those which hold solely by virtue of meaning, or which are knowable solely by virtue of meaning. An example of a truth that would generally be held by defenders of analytic truth to be analytic is that all bachelors are unmarried; an example of a truth that would generally be held not to be analytic is that most crows are black. In a moment, I will explain some of the controversy surrounding analyticity. First, however, I will draw on Lewis's (1989) work to develop a model of conceptual analysis that purports to yield analyses that state analytic truths, so understood.

Lewis (1989) argues, roughly, that something is valuable iff (Def) we would be disposed, under conditions of full imaginative acquaintance, to desire to desire it. Lewis

[4] I should emphasise that these models are developments of ideas in the work of these philosophers, rather than serious attempts at exegesis.

defends this as an analytic truth, but an analytic truth that is ‘unobvious’ and ‘equivocal’. It is unobvious in the sense that it is possible for someone intelligibly to express genuine doubt about whether it is extensionally correct (130). Lewis explains the notion of equivocal analyticity as follows:

> Something may be analytic under one disambiguation but not another, or under one precisification but not another… …If differing versions of a concept (or, if you like, different but very similar concepts) are in circulation under the same name, we will get equivocal analyticity. It is analytic under one disambiguation of ‘dog’ that all dogs are male… …It is analytic under some precisifications of ‘mountain’ that no mountain is less than one kilometre high… …I suggest that the dispositional theory of value, in the version I have put forward, is equivocally as well as unobviously analytic. I do not claim to have captured the one precise sense that the word ‘value’ bears in the pure speech, uncorrupted by philosophy, that is heard on the Clapham omnibus… …I take it, rather, that the word ‘value’, like many others, exhibits both semantic variation and semantic indecision. The best I can hope for is that my dispositional theory lands somewhere near the middle of the range of variation and indecision – and also gives us something that I, and many more besides, could be content to adopt as our official definition of the word ‘value’, in the unlikely event that we needed an official definition. (Lewis 1989: 130-131)

The core idea here is that, insofar as a term exhibits both semantic variation and semantic indecision, we should expect any relatively precise philosophical analysis of a concept expressed by that term to be analytic only under some disambiguation and some precisification of that term. By ‘semantic variation’, Lewis means the phenomenon whereby some words are polysemous between highly similar senses, such that these words can be used to express subtly different concepts in different contexts of utterance. By ‘semantic indecision’, Lewis means

the phenomenon that occurs when 'we have not troubled to settle which of some range of precise meanings our words are meant to express' (1986: 244, n.32). ('Mountain' is his example of a term exhibiting semantic indecision).

Now, insofar as any relatively precise philosophical analysis of a concept expressed by a certain term – such as 'morally wrong' – is analytic only under some disambiguation and some precisification of that term, then our reasons for *privileging* some such analysis must be largely evaluative: it is because the concept captured by that analysis is especially valuable that we have reasons for privileging it as the correct philosophical analysis (or, at least, the correct philosophical analysis relative to a certain evaluative interest – more on this below). Lewis's hope that his dispositional analysis 'gives us something that I, and many more besides, could be content to adopt as our official definition of the word 'value'' seems to bear witness to this point (1989: 131; cf. also 136). We can, then, find in Lewis's work a model of conceptual analysis that occupies the space between description and prescription that MB aspires to occupy. The model is descriptive insofar as conceptual analyses, as it understands them, are meant to land somewhere near the middle of the range of semantic variation and indecision exhibited by the relevant terms. And it is prescriptive or ameliorative insofar as, within this range, our reasons for privileging some particular analysis are evaluative. The resultant analyses can therefore be understood partly as proposals for what concepts we *should* use. Call this the 'Lewisian model'.[5]

The Lewisian model is one model of conceptual analysis that is consistent with my defence of MB, but it is not the only model. Scepticism about analytic truth has been an

[5] There is a further aspect of Lewis's (1989) understanding of conceptual analysis that is worth mentioning briefly, even though it is not relevant to my defence of MB. Lewis suggests that sometimes there may not be a 'perfect deserver' of a given term (136). What he means is that in some cases there may be nothing answering to all of the platitudes associated with a given term. But there may be various 'imperfect deservers': things that do not answer to all of these platitudes, but answer to enough of them to deserve to go by the name of the term in question. Lewis claims that this illustrates one way in which an analysis may be equivocally analytic: it may capture one imperfect deserver of the term in question, even if it is not the only imperfect deserver.

important source of scepticism about the method of conceptual analysis (Laurence and Margolis 2003: 254-255). If conceptual analyses purport to yield analytic truths, and there are no such truths to be had, or at least no philosophically interesting ones, then conceptual analysis as a philosophical methodology must be misguided. In a moment, I will draw on Gibbard (1990) to develop a different model of conceptual analysis that does not purport to yield analytic truths. Given that my defence of MB is compatible with either the Lewisian model developed above or the Gibbardian model developed below, MB cannot be undermined by doubts about analytic truth. First, however, let me explain one of the main considerations that has led many philosophers to reject the claim that there are any philosophically interesting analytic truths.

The consideration in question, which was influentially brought to bear against the notion of analytic truth by W. V. O. Quine (1951), is the claim that confirmation is holistic: individual statements are never confirmed in isolation, but only as parts of overall theories. In Quine's view, this has the upshot that any individual statement, taken on its own, could in principle be rejected, provided that enough changes are made in your wider theory or views to accommodate its rejection (1951: 40). An example of Hilary Putnam's (1962) provides a vivid illustration of this thought with respect to the statement 'All cats are animals'. Putnam invites us to imagine discovering that all of the cats there have ever been were in fact robots built by Martians (660-661). Putnam suggests that this discovery would lead competent users of 'cat' and 'animal' to conclude, not that there have never been any cats, but rather that cats are not animals (660). So, even an apparently analytic statement such as 'All cats are animals' turns out to be in principle revisable provided that enough changes in your wider views are made to accommodate its rejection.

The problem this poses for the notion of analytic truth is that statements of analytic truths are precisely meant *not* to be revisable: if analytic truths are true in virtue of meaning

alone, then to give up an analytic truth is to change the meanings of the terms used to state it. So, if any individual statement, taken on its own, could in principle be rejected without changing the meanings of the terms used to state it, then there are no analytic truths. Note, moreover, that to make trouble for conceptual analysis as a philosophical methodology that purports to yield analytic truths, a considerably weaker claim will do. Even if there are some trivial analytic truths – perhaps 'All bachelors are unmarried' supplies an example – it may be that any individual statement concerned with the concepts that have most interested philosophers – KNOWLEDGE, CAUSE, and MORALLY WRONG, for instance – is in principle revisable (cf. Putnam 1975).

I will not take a stand on whether this argument, or any other argument, shows that there are no philosophically interesting analytic truths. I do, however, want my defence of MB to be consistent with this possibility. To this end, I will draw on Gibbard (1990) to develop a model of conceptual analysis that does not purport to yield analytic truths. In the course of explaining how conceptual analysts might proceed if it is granted that there is not a philosophically significant analytic/synthetic distinction, Gibbard writes:

> An analysis must be judged by its fruits: How much does it explain? How much thought does it make intelligible, and how little would it have us dismiss as unintelligible? How good an explanation emerges of the role of a kind of language in human life? How, by these tests, do the alternatives compare?... …An analysis can be offered not as a bald statement of fact about what people mean, but as a proposal. Where a term is problematical, a new and clearer sense may serve its purposes – or some of them… …Any philosophical analysis strains its concept. We can learn about a concept by seeing what choice of strains it offers. When an analysis keeps us from saying things we want to say, then we have to think how important it is to go on saying them, and we have to think about costs. (1990: 32; cf. also 1992: 199-200)

In this passage, Gibbard sketches a model of conceptual analysis that occupies the space between description and prescription that MB aspires to occupy. On the one hand, it is a desideratum for analyses to fall broadly in line with the concepts we in fact use – hence the concern with how much of our current thought an analysis makes intelligible, and how far it strains its concept. But on the other hand, it is a desideratum for analyses to yield concepts that are worth keeping, insofar as they serve important purposes. In Gibbard's view, this understanding of conceptual analysis does not presuppose that a philosophically interesting analytic/synthetic distinction can be drawn, and hence does not purport to yield analyses that state analytic truths (1990: 33). This means, I take it, that particular concerns falling under the descriptive desideratum have a defeasible status. With respect to any aspect of our current use or understanding of a concept, there is no guarantee, prior to analysis, that it will end up preserved in the final analysis. If it conflicts with other aspects of our current use, or with the valuable purposes we want our concept to serve, then it may be discarded. Call the model of conceptual analysis outlined here the 'Gibbardian model'.

In a moment, I will explain why the Lewisian and Gibbardian models of conceptual analysis are fruitful models, and defend them against an important objection. First, however, we should note that despite the fact that these models differ with respect to the issue of analytic truth, there is a striking convergence in the kinds of considerations that are relevant to defending them. Both models of conceptual analysis aim to yield analyses that are broadly in line with central features of how we use and understand our current concepts. As I will put it, both aim to respect for the most part the judgmental and inferential dispositions that are typically had by those who possess the relevant concepts. I will use 'platitudes' as a term of art for statements of these dispositions (I say more about how I understand platitudes in Chapter

2, and what it means for an analysis to 'respect' them).[6] The models differ as to the status of these platitudes. On the Lewisian model, these platitudes are such that giving them up would involve changing the meanings of the relevant terms – at least with respect to some disambiguations and precisifications of them. In contrast, on the Gibbardian model, each of the platitudes, taken individually, is such that you could in principle reject it without changing the meaning of the relevant term or the identity of the concept it expresses. Just as both models involve a descriptive desideratum, so too do both models involve a prescriptive or ameliorative desideratum. The analyses in question should yield concepts that are worth having, relative to some evaluative interest.[7]

My defence of MB will remain neutral between the Lewisian and Gibbardian models, and hence neutral on whether there are any philosophically significant analytic truths. Chapter 2 argues that MB respects a central range of platitudes associated with our deontic moral concepts. If this is right, then MB meets the descriptive desideratum at play in the Lewisian and Gibbardian models. Chapter 3 argues that the account of deontic moral concepts provided by MB explains how deontic moral concepts serve valuable social functions, such that we would be considerably worse off without these concepts. If this is right, then MB meets the prescriptive desideratum at play in both models.

Let me say more about the prescriptive or ameliorative aspect of my defence of MB. That a conceptual analysis is one we should accept, given a certain evaluative interest, would hold little claim to our attention if the value itself were unimportant. In defending MB, I will argue that deontic moral concepts, as MB understands them, are valuable concepts from the

[6] Smith (1994: 30) uses 'platitudes' in a similar way, although his understanding of conceptual analysis places less emphasis on amelioration than mine.

[7] Neither of these models, as far as I can see, is essentially tied to a particular view of the ontology of concepts – for instance, to the view that concepts are mental representations, or the view that they are abstract entities (see Laurence and Margolis (1999: 5-8) for discussion of these views). Accordingly, my defence of MB will remain neutral on this issue.

perspective of not just any values, but vitally important ones. To anticipate, I will argue that deontic moral concepts serve two valuable social functions. First, they serve 'the Deliberative Reliability Function', by which I mean the function of protecting important interests through encouraging reliable patterns of deliberation and motivation. Examples of these interests include the interests we have in not being killed, hurt, or lied to (and the further interests we have in being able to rely on not being treated in these ways). And second, deontic moral concepts serve 'the Respect Function', by which I mean the function of providing social recognition of the dignity of persons and their worthiness of respect.

I have frequently talked about the value of concepts. Although I will continue to talk in this way, what I ultimately have in mind is the value of people *using* certain concepts in thought and talk. Adrian Moore helpfully distinguishes between two ways of using concepts: 'disengaged use' and 'engaged use' (2006: 137; cf. also Williams 1986: 203-204). A disengaged user of a concept is able to understand what it applies to, understand others when they apply it, and understand the kinds of inferences it is thought to licence. An engaged user of a concept can not only do these things, but also *lives by* the concept: they use it in how they think about the world and conduct their affairs. Moore illustrates this distinction with the concept SABBATH (Ibid.). Non-Jewish people can use this concept in a disengaged way, but only observant Jews use it in the engaged way. In talking about the value of deontic moral concepts, I am interested in the value of engaged use of these concepts.[8] Engaged use of deontic moral concepts brings with it a host of (defeasible) inferential and judgmental dispositions: dispositions to morally blame those you judge to have acted morally wrongly without a moral excuse, for example, and dispositions to judge that it is appropriate for unexcused moral wrongdoers to make amends through such things as apologies, offers of compensation, and

[8] If some version of 'moral judgment internalism' is correct, and there is a necessary or internal connection between deontic moral judgments and motivation, then perhaps it is not even possible to use deontic moral concepts in the disengaged way.

attempts at re-form. The value of engaged use of deontic moral concepts stems from the value of people having these dispositions (and also having communicative devices for co-ordinating these dispositions).

Why should we be interested in MB? There are at least three reasons. First, MB puts us in a position to respond to philosophers who argue that deontic moral concepts are pernicious concepts that we would be better off without. Perhaps the best-known representative of this view is G. E. M. Anscombe, who argues that deontic moral concepts are 'survivals, or derivatives from survivals, from an earlier conception of ethics [a divine law conception] which no longer generally survives, and are only harmful without it' (1958: 1).[9] Insofar as deontic moral concepts, as understood by MB, serve valuable social functions (and are non-religious), we can appeal to MB to respond to this view. Note that this requires that MB meets both the descriptive and prescriptive desiderata at play in the Lewisian and Gibbardian models. MB needs to meet the prescriptive desideratum to show that deontic moral concepts, as it understands them, are indeed valuable concepts. But MB also needs to meet the descriptive desideratum on pain of changing the subject: if deontic moral concepts, as MB understands them, are not broadly in line with the deontic moral concepts we in fact use, then a defender of MB would actually be in partial *agreement* with Anscombe: they would think we would be better off jettisoning our actual deontic moral concepts and (they would add) replacing them with importantly different concepts.[10]

[9] Another philosopher who might be thought to fall into this category is Bernard Williams, especially in his (1985). Williams's views are more nuanced and qualified than Anscombe's, however. Although Williams argues that we would be better off without the concept of moral obligation, he uses 'moral obligation' as a term of art for a particularly demanding conception of obligation that he takes to be widespread in modern thought (1985: 174-196). At the same time, Williams argues that argues that the concept of obligation, when freed from this demanding conception of it, is a valuable one. When he explains what is distinctive about the concept of obligation, he often appeals to its connection with blame (1985: 186-188). It is, I think, accurate to claim that Williams does not think we would be better off without deontic moral concepts *as MB understands them* entirely, but only that we would be better off without a particularly demanding conception of them.

[10] This assumes that users of ameliorated concepts can be in substantive disagreement with users of unameliorated concepts, provided the ameliorated and unameliorated concepts are sufficiently similar. For discussion and

A second reason why we should be interested in MB is that it can help us to proceed more effectively in our first-order moral inquiries, by giving us a clear account of what we are doing when we engage in these inquiries and preventing us from talking at cross-purposes (Gibbard 1990: 33; 1992: 199-200). Admittedly, this point needs to be qualified by recognising that conceptual analyses (on both the Lewisian and Gibbardian models) are subject to a neutrality desideratum: defenders of such analyses should typically be wary of taking on first-order commitments, since they must show that these commitments can be established on the basis of the platitudes associated with deontic moral concepts. (I explain and motivate this desideratum in Chapter 4). However, this does not mean that MB cannot assist us in our first-order moral inquiries at all. Indeed, I argue in Chapter 4 that MB can help us to make important progress in assessing consequentialist moral theories.

A third and final reason why we should be interested in MB is that it can help us to understand the relations between deontic moral concepts and other normative concepts that we use. In Chapter 7, I take up the question of the relation between moral wrongness and normative reasons in light of MB. I argue that MB supports what I call 'the Normativity Thesis'. According to the Normativity Thesis, if it is morally wrong for an agent to φ, then that agent has strong normative reasons not to φ. However, whether these reasons are decisive or even sufficient is left open by the Normativity Thesis. I argue further that MB does not support any stronger theses concerning the connection between moral wrongness and normative reasons, such as the Overridingness Thesis, according to which we always have decisive normative reasons not to act morally wrongly.

Before moving on to explain how I understand metaphysical analysis, I will consider an important objection to the method of conceptual analysis. Earlier, I explained how my

---

defence of this assumption (albeit framed primarily at the level of *terms* rather than *concepts*), see Cappelen (2018: 107-121; 2020: 140-141).

defence of MB is consistent with either the claim that there are philosophically interesting analytic truths or its rejection. Doubts about analytic truth have been one major source of scepticism about the value of the method of conceptual analysis in meta-normative theory. Another has been the 'Open Question Argument'.[11] The Open Question Argument was initially deployed against attempts to give naturalistic analyses of normative concepts, but a version of it can be deployed against analyses like MB, which try to analyse normative concepts in terms of further, allegedly more basic normative concepts. The argument works by inviting us to ask whether the left-hand side of a given conceptual analysis always aligns with its right-hand side. In the case of MB, the relevant question is, 'Is it always morally wrong to perform actions[12] that violate standards such that, if you violated those standards without a moral excuse, you would be morally blameworthy for violating them?'. The next step in the argument is to urge that this question is 'open' – that is, such that someone could sincerely ask it as a genuine question (as opposed, say, to a joke) without thereby displaying linguistic, conceptual, or logical confusion. The final step in the argument is to claim that if this question is open, then MB cannot provide the correct analysis of MORALLY WRONG. More formally, the argument can be presented as follows:

> P1: The question, 'Is it always morally wrong to perform actions that violate standards such that, if you violated those standards without a moral excuse, you would be morally blameworthy for violating them?', is open.
>
> P2: If P1, then MB does not provide the correct analysis of MORALLY WRONG.
>
> (From P1 and P2) C: MB does not provide the correct analysis of MORALLY WRONG.[13]

[11] This argument was first presented by Moore (1903) and has since been extensively discussed. For overviews, see Darwall et al. (1992), Miller (2013: Chapter 2), and McPherson (2013).

[12] This is a slight oversimplification, since MB leaves it open what are the possible objects of deontic moral assessment.

[13] A note on my use of 'P1', 'P2', 'C' etc.: each chapter will use this notation in a standalone way, such that, e.g., P1 in the Introduction may not be the same as P1 in Chapter 1.

I will not question P1 – on the contrary, it seems correct. However, I will argue that P2 is false. Both the Lewisian and Gibbardian models have ample resources for explaining why the question stated in P1 is open.

One resource for explaining the openness of the question stated in P1 that is available on the Lewisian model is to appeal to equivocal analyticity (cf. Lewis 1989: 130). Insofar as MB is analytic only under some disambiguation and precisification of 'morally wrong', we should expect it to be possible to question sincerely whether its right-hand side always coincides with its left-hand side without thereby displaying linguistic or logical confusion. This is because there will be some senses of 'morally wrong' in circulation on which the equivalence postulated by MB is not analytic, and indeed not even true. On the Lewisian model, MB's claim to provide the correct philosophical analysis of MORALLY WRONG, relative to some evaluative interest, is not that it uniquely captures the *only* concept expressed by uses of 'morally wrong', but rather that it captures an especially valuable one.

There may be further resources available on the Lewisian model for explaining open questions. I suggested that the basic inputs for conceptual analyses on this model are the 'platitudes' surrounding the concepts to be analysed: these are statements of the inferential and judgmental dispositions typically had by those who possess the concepts in question. It is not clear, however, that it is a condition on possessing these dispositions that one must know that one possesses them, or know what, in light of these dispositions, are the best analyses of the relevant concepts (cf. Smith 1994: 38). As Michael Smith writes:

> Why are analyses unobvious and informative? Because even though someone who has mastery of some concept C must *have* certain inferential and judgmental dispositions, it may not be transparent to her what these inferential and judgmental dispositions are, and so, *a fortiori*, it need not be transparent to her what the best summary or

> systematisation of the platitudes that describes these dispositions is. Whereas mastery of a concept requires knowledge-how, knowledge of an analysis of a mastered concept requires us to have knowledge-that about our knowledge-how. [emphasis in original] (1994: 38)

Note, moreover, that given that the Lewisian model involves a prescriptive desideratum, a further gap opens up between knowledge of the platitudes surrounding the relevant concepts and knowledge of the best analyses of these concepts: this latter kind of knowledge requires knowledge of the values that are furthered by the relevant concepts. It seems, then, that the Lewisian model leaves us with plenty of resources for explaining the openness of the question stated in P1.

The Gibbardian model also has resources for explaining the openness of this question. Indeed, since it does not purport to yield analyses that state analytic truths (i.e., truths which hold solely by virtue of meaning, or which are knowable solely by virtue of meaning), there is not even a *prima facie* expectation that the question stated in P1 should not be open. Knowledge that MB is the correct analysis of MORALLY WRONG, on the Gibbardian model, involves not only knowing that MORALLY WRONG, as understood by MB, respects central aspects of how we ordinarily use and understand this concept, but also knowing that it is a valuable concept worth keeping. Clearly, it is possible to fail to have such knowledge without exhibiting linguistic or logical confusion. Hence, the Gibbardian model predicts that it should be possible sincerely to ask, 'Is it always morally wrong to perform actions that violate standards such that, if you violated those standards without a moral excuse, you would be morally blameworthy for violating them?', without thereby betraying linguistic or logical confusion, even if MB supplies the best analysis of MORALLY WRONG.

Let me make a final remark about my defence of MB. My aim is to develop new arguments for MB and defend it against objections. However, although I will briefly discuss and criticise some rivals to MB, I will not undertake a systematic comparison of the merits of MB with the merits of these rivals.[14]

1.1.2 Metaphysical Analysis

As well as defending MB as a conceptual analysis, I also defend MB as a metaphysical analysis (or, equivalently, a *real* definition): that is, an analysis of the moral wrongness *relation*. By an analysis of a property/relation, I mean an account of *what it is* for something to be F, where 'F' picks out a property or relation. Examples of such claims are commonplace in philosophy and other disciplines. Some examples include: for something to be water is for it to be H2O; for something to be a square is for it to be an equilateral four-sided figure; and for something to be good is for there to be sufficient reasons to desire it.

There are different ways of understanding what metaphysical analyses amount to. One view is that metaphysical analyses, insofar as they are successful, explain how complex properties/relations (e.g., being square) reduce to structured simpler properties/relations (e.g., being four-sided, being equilateral, and the conjunction structure) (Schroeder 2007: 67-72). Another view, which is not necessarily in competition with the first, is that metaphysical analyses should aim to state the essences of the properties or relations under analysis (Fine 1994; Wedgwood 2007: 136-144). And a third view, which, again, is not necessarily in competition with the previous two, is that metaphysical analyses, if they are successful, explain

[14] Indeed, my defence of MB, as far as it goes, is consistent with the view that there are multiple moral wrongness concepts that are valuable to moral thought in different ways (cf. Parfit 2011). Someone who takes this view could see my defence of MB as an attempt to analyse one of these concepts and explain its distinctive importance.

how complex properties/relations reduce to the simpler properties/relations that ground these complex properties/relations (Rosen 2015a).

My defence of MB as a metaphysical analysis is consistent with any of these views. However, I will make a few important assumptions about metaphysical analysis. The first is that, if F is metaphysically analysable in terms of G, then when something is G, it cannot be G *because* it is F. For example, if to be a square is to be an equilateral four-sided figure, then something cannot be an equilateral four-sided figure because it is a square. The second assumption is that metaphysical analyses are not ontologically committing: a true claim of the form, 'For something to be F is for it to be G', does not entail that there are any Fs. For this reason, MB as a metaphysical analysis is compatible with moral error theory. The third assumption about metaphysical analysis I make is that the kind of explanation that a metaphysical analysis of moral wrongness provides is, in principle, different from the kind of explanation provided by moral theories such as consequentialism, contractualism, and Rossian pluralism. Although these theories can be defended as claims about what it is to be morally wrong, a different (and more common) way of defending them is as claims about which actions are morally wrong and why. So understood, a metaphysical analysis of moral wrongness could in principle be paired with any of a range of first-order moral theories. In the case of MB, disagreement between rival moral theories can be understood as disagreement about which standards are such that, if someone violated them without a moral excuse, they would be morally blameworthy for violating them. (For the sake of readability, I will sometimes refer to these standards as 'moral blameworthiness-related' standards. By a 'standard', I mean a criterion by which one can, in principle, judge or appraise various things on the basis of meeting or failing to meet the condition set by the criterion (cf. Copp 1995: 19). I say more about how I understand standards in Chapter 2.)

None of this is to say that the projects of defending a metaphysical analysis of moral wrongness and defending a first-order moral theory are entirely separate from one another. Indeed, we will see in Chapter 4 that combining MB with some common versions of consequentialism yields implausible combinations of views, given the implications these combinations of views have for claims about the conditions under which agents are morally blameworthy. However, while there are some interesting points of contact between MB and first-order moral theorising, MB, as understood here, has different explanatory ambitions insofar as it aims to give an account of what it is to be morally wrong.

MB leads to a particular account of the structure of the deontic moral domain (cf. Skorupski 2010: 318-320). We can distinguish between *the fact that it is morally wrong for an agent to φ* and *the facts that make it morally wrong for an agent to φ*. Call these latter facts 'moral wrong-makers'. Examples of moral wrong-makers might include the fact that φ-ing involves breaking a promise, or the fact that φ-ing involves causing great pain. According to MB (read as a metaphysical analysis), for it to be morally wrong for an agent to φ is for φ-ing to violate moral blameworthiness-related standards. A fact is a moral wrong-maker, on this view, if and only if it makes it the case that φ-ing violates moral blameworthiness-related standards. (Facts, such as the fact that φ-ing involves breaking a promise, make it the case that objects of deontic moral assessment violate standards by making it the case that these objects fail to meet the conditions set by these standards).

According to the account of the structure of the deontic moral domain supported by MB, then, we can give a full explanation of why some object of assessment, φ, violates moral blameworthiness-related standards by citing such facts as: the fact that φ-ing involves breaking a promise, that the promise is uncoerced and not immoral, that no comparable good will be brought about by breaking the promise, and so on. We do not need to cite a further fact about

moral wrongness to provide a full explanation of why φ-ing violates moral blameworthiness-related standards.

This account of the structure of the deontic moral domain can be contrasted with some different accounts that reject MB as a metaphysical analysis. Provided that MB is extensionally correct – and I argue that it is in Chapter 4 – it seems that there is one main family of rival views concerning the structure of the deontic moral domain.[15] On the views I have in mind, we *do* need to cite facts about moral wrongness to explain facts about moral blameworthiness. More precisely, the moral wrongness relation is a distinct relation (which may or may not be analysable), and facts about moral wrongness explain facts about whether φ-ing violates moral blameworthiness-related standards. We saw earlier that, if F is metaphysically analysable in terms of G, then when something is G, it cannot be G because it is F. Given this, the claim that facts about moral wrongness explain facts about whether φ-ing violates moral blameworthiness-related standards is inconsistent with MB.

MB as a metaphysical analysis does not simply follow from MB as a conceptual analysis. To be sure, there are some analyses that succeed both as conceptual and metaphysical analyses. For example, 'X is Y's aunt if and only if (Def) X is a woman sibling or sibling-in-law of one of Y's parents', plausibly supplies the correct analysis of the concept AUNT and the property of being an aunt. But conceptual and metaphysical analyses may not always coincide in this way. For example, it seems entirely possible to hold that, 'Something is red if and only if it is such as to look red to normal observers under standard conditions', supplies the correct analysis of the concept RED while denying that it gives the correct analysis of the property of being red (cf. Pettit 1991, 1998). The property of being red, we might hold, is not this

[15] Another rival family of views is that moral wrongness and moral blameworthiness are in some way mutually interdependent, with neither relation being metaphysically more basic than the other. I discuss this kind of view briefly in 1.2.1.

dispositional property but rather the objective property or properties that realise this disposition (e.g., a reflectance property of surfaces) (cf. Jackson and Pettit 2002). Analogously, it seems entirely possible for someone to accept MB as a conceptual analysis while claiming that the moral wrongness relation is not to be analysed in terms of moral blameworthiness, but is rather a distinct relation that explains moral blameworthiness. Given this, we cannot reasonably defend MB as a metaphysical analysis by simply redeploying our arguments for it as a conceptual analysis.

My defence of MB as a metaphysical analysis will be sensitive to this point, and accordingly I will argue in Chapter 2 that there are good reasons for accepting MB not just as a conceptual analysis but also as a metaphysical analysis. I will argue that MB gains plausibility as a metaphysical analysis in virtue of conforming to an attractive general pattern of analysis of different kinds of requirements, permissibility, and wrongness across different normative domains (for ease of presentation, I focus in particular on requirements). In this way, MB forms part of an attractive explanation of what unifies different kinds of requirements across different normative domains.

1.2 Sentimentalism

At the most general level, sentimentalism in meta-normative theory holds that emotions play an essential role or roles in normative talk, thought, and/or reality (I use 'normative' as a catch-all for the deontic and the evaluative). For example, a sentimentalist might hold that certain normative concepts or properties (e.g., funniness) are essentially related to emotions (e.g., amusement), that emotions play a crucial role in the explanation of the meaning of certain normative statements, or that emotions afford our primary mode of access to knowledge of certain kinds of normative truths.

Sentimentalism, so understood, encompasses a wide and varied range of views. I will not attempt to map out the relations of MB to all of the possible views in this range. Instead, I will focus primarily on the subfamily of sentimentalist views to which MB belongs: what are sometimes called 'neo-sentimentalist' analyses (McDowell 1985; Wiggins 1987; Gibbard 1990; D'Arms and Jacobson 2000a, 2003, 2017, 2022, 2023; Skorupski 2010; Tappolet 2011, 2016; Kauppinen 2017; Shoemaker 2017).[16] Neo-sentimentalists analyse certain normative concepts and/or properties/relations[17] in terms of the *fittingness* of certain emotions. In 1.2.1, I introduce neo-sentimentalism and clarify the fittingness relation. In 1.2.2, I discuss two arguments for neo-sentimentalism. Finally, in 1.2.3, I consider two well-known, general difficulties with neo-sentimentalist analyses – 'the wrong kind of reasons problem' and 'the distance problem'.

1.2.1 Neo-Sentimentalism

Neo-sentimentalists can target a narrower or wider range of normative concepts and/or properties. Focussing for now on concepts, some normative concepts, such as DISGUSTING, SHAMEFUL, and ADMIRABLE, seem especially suited to a neo-sentimentalist treatment (D'Arms 2005: 2; Tappolet 2016: 81). Despite the interest of these normative concepts, more ambitious versions of neo-sentimentalism target a wider range of normative concepts that are of more central concern to normative theorists, such as MORALLY WRONG (Gibbard 1990; Skorupski 2010), VIRTUOUS (McDowell 1985), and WELL-BEING (Rossi and Tappolet 2022). MB is a contribution to this more ambitious neo-sentimentalist project.

[16] The term 'neo-sentimentalism' was coined by D'Arms and Jacobson (2000a). In labelling this family of views 'neo-sentimentalist', they meant, of course, to contrast it with some earlier sentimentalist approaches. I explain the intended contrast (and voice some misgivings about it) in a moment.
[17] In the interests of readability, I will tend to drop '/relations'.

An important divide among neo-sentimentalists concerns whether they defend non-circular analyses of normative concepts and/or properties (Gibbard 1990; Skorupski 2010; D'Arms and Jacobson 2017), or else defend avowedly circular elucidations of normative concepts and/or properties (McDowell 1985; Wiggins 1987; Tappolet 2016). My aim is to defend non-circular conceptual and metaphysical analyses of moral wrongness, permissibility, and requirement. Accordingly, in Chapter 5 I defend accounts of the nature of moral blaming emotions and what it is for them to be fitting that make no essential reference to deontic moral concepts and relations, and hence are suited to figuring in non-circular analyses of these concepts and relations.

Non-circular neo-sentimentalist analyses of normative concepts and relations, such as MB, are largely neutral with respect to broader meta-ethical questions about the nature of normative talk, thought, and reality. MB is neutral on whether fittingness judgments are fundamentally cognitive or non-cognitive in nature (or whether they are hybrid states of some kind), and hence is compatible with cognitivist, non-cognitivist, or hybrid views concerning the nature of moral judgment. MB is neutral on whether the moral wrongness relation is a natural or non-natural relation, and on whether this relation is ever instantiated. Hence, MB could be paired with non-naturalist or naturalist versions of moral realism, and also with moral error theory. And insofar as quasi-realists claim that there are moral properties and relations, it might be possible for MB, even as a metaphysical analysis, to be paired with various forms of moral anti-realism.

A further issue on which neo-sentimentalists can be to some extent flexible concerns the nature of emotional fittingness. Fittingness has been much-discussed in recent years.[18]

[18] See, e.g., D'Arms and Jacobson (2000a, 2000b); Rabinowicz and Rønnow-Rasmussen (2004); Svavarsdóttir (2014); McHugh and Way (2016, 2022); Tappolet (2016); Howard (2018, 2019); Rowland (2019); Naar (2021); Berker (2022); and D'Arms (2022).

Some philosophers argue that it is a normatively primitive relation, while others argue that fittingness can be analysed in terms of other categories such as accurate representation or reasons or values. As we will see in Chapter 5, 'pure' accurate representation views – views according to which emotional fittingness is *solely* a matter of accurate representation – are difficult to square with neo-sentimentalist analyses that aim to be non-circular. However, neo-sentimentalist views could, in principle, be paired with a wide range of alternative views of fittingness. For example, the views that fittingness is normatively primitive, or analysable in terms of reasons or values, could all, in principle, be paired with neo-sentimentalist analyses (provided that the analyses of fittingness in question do not appeal to the normative concepts or properties that the neo-sentimentalist seeks to analyse).

Although I will make some suggestions as to how fittingness is best understood, my defence of MB will remain largely neutral between the views that fittingness is normatively primitive, or else analysable in terms of further normative categories such as reasons or values. Now, it might be worried that leaving the issue of how fittingness is best understood open in this way means that MB is not very informative. MB, it might be thought, attempts to analyse deontic moral concepts and relations in terms of a concept and relation that, absent further explication, is at least as obscure. To address this concern, I will now clarify fittingness by comparing and contrasting it with some other kinds of normative assessment. While this will not amount to an analysis of fittingness, the desired result is that the fittingness relation will emerge as one on which we have a sufficiently clear grasp that we can usefully put it to substantive work in meta-normative theorising, even if we remain neutral on whether it is normatively primitive.

The fittingness relation is the relation that holds between an object, a subject, and a response when the object merits – or, equivalently, is worthy of – that response from that subject. Many terms for responses have associated normative terms formed by suffixation, such

that, roughly, the response in question is fitting if and only if its object falls under its associated normative term.[19] Examples of such terms include 'shameful', 'admirable', 'amusing', and 'contemptible'. So, for example, it would be fitting to admire a great humanitarian for their altruism – being altruistic is admirable. But it would not be fitting to admire a ruthless dictator for their cruelty – being cruel is not admirable.[20]

Assessing responses for fit is not the only way of normatively assessing responses (D'Arms and Jacobson 2000b). For instance, an emotion may be prudentially good even though it is not fitting, or prudentially bad even though it is fitting. If an evil demon credibly threatens to torture you unless you admire something that is not admirable, feel ashamed of something that is not shameful, or feel amused by something that is not amusing, then while it would be prudentially good for you to feel these emotions they would not be fitting (cf. Rabinowicz and Rønnow-Rasmussen 2004: 407-408). (I leave to one side for now the controversial question of whether the demon's credible threat is a *reason* for these emotions). Another form of normative assessment of responses that is different from fittingness assessment is moral assessment. For example, it may be morally bad for a recently widowed parent to feel all the grief it would be fitting for them to feel, because this would risk further harm to their children (cf. D'Arms and Jacobson 2000b: 77). Note that the claim here is that moral and prudential assessment *of responses* is different from fittingness assessment of responses. This is consistent with thinking that moral and prudential assessment of the *objects* of responses is relevant to fittingness

[19] There are a few reasons why 'roughly' is needed. First, fittingness involves a proportionality condition: to be fitting, the strength of a response must be proportional to the value of its object (cf. D'Arms and Jacobson 2000b: 73-74). For example, very strong shame towards something that is only a little shameful would not be fitting. Moreover, the fittingness of a response often depends on contextual factors: it would not be fitting for me, now, to fear a dangerous cliff while I am sat at my desk (cf. D'Arms and Jacobson 2023: 72). Rather, fear would only be fitting for people who are close to its edge. I discuss the connection between fitting emotions and properties such as shamefulness, admirableness, amusingness, and so on further in the next section.

[20] For a similar initial gloss on the fittingness relation, see Howard (2018: 1-2).

assessment of responses. For example, someone may be admirable partly in virtue of their moral qualities.

Another kind of normative assessment with which fittingness assessments can usefully be contrasted are generic deontic assessments: assessments of what we are normatively required to do (i.e., required *simpliciter* to do), normatively permitted to do (i.e., permitted *simpliciter* to do), and so on.[21] In many, if not all cases, claiming that a response is one that we are normatively required to have is intuitively stronger than claiming that it would be fitting or merited (cf. Svavarsdóttir (2014: 98-101); Berker (2022: 36-40); McHugh and Way (2022: 75-76)). For example, it is one thing to say that a funny joke merits, or is worthy of, amusement, but it seems importantly different, and stronger, to say that we are required to be amused by it. On the other hand, claiming that a response is one that we are normatively permitted to have is typically, if not always, intuitively weaker than claiming that it is fitting: fittingness is positive in a way that mere permission is not (cf. Ibid.). For example, saying that something merits, or is worthy of, shame, seems to commend shame in a way that merely saying that shame is permissible does not. In making these brief remarks, I do not take myself to have shown conclusively that fittingness cannot be analysed in terms of normative requirement or permissibility. [22] However, contrasting fittingness assessments with generic deontic assessments in these ways does help to bring out that we have a strong, intuitive grip on the fittingness relation.

Before moving on to discuss some motivations for defending neo-sentimentalist analyses, let me first say something about the label 'neo-sentimentalism' and the contrast with earlier sentimentalist approaches to which it alludes. In their influential paper 'Sentiment and

[21] It is controversial whether good sense can be made of generic deontic assessments. For doubts, see, e.g., Copp (1997). For defence, see, e.g., Dorsey (2016); McPherson (2017). I will assume for the purposes of clarifying the fittingness relation that good sense can be made of generic deontic assessments, but if this assumption is rejected the clarifications of fittingness given above still stand.

[22] See Berker (2022: 37-40) for critical discussion of some sophisticated proposals.

Value', Justin D'Arms and Daniel Jacobson contrast *neo*-sentimentalist analyses of normative concepts and properties, which focus on the fittingness of certain emotional responses, with earlier versions of sentimentalism that allegedly do not give a role to fittingness (2000a: 724-729). Chief among these earlier versions are *dispositionalist* accounts, which attempt to analyse normative concepts and properties in terms of how certain, possibly idealised, individuals or groups would be disposed to respond emotionally under certain, possibly idealised, circumstances (726-729). D'Arms and Jacobson argue that dispositionalist accounts struggle to capture adequately the normative force of the concepts under analysis (727). Concerning a simple version of dispositionalism that focuses on the responses of statistically normal people, they write:

> The central normative role of concepts like shameful, funny, and enviable is to govern the associated sentiments, but a dispositional account of these concepts would prevent them from playing that role in the wide range of cases in which one wants to contest popular opinion. (2000a: 727)

As D'Arms and Jacobson recognise, more sophisticated versions of dispositionalism may try to avoid this problem by focussing on the responses of idealised individuals or groups, rather than statistically normal people. But they worry that such accounts will still struggle to capture the normative force of the concepts under analysis (728). Moreover, such accounts threaten to be empty or tautologous, if the idealised judge is identified, in effect, as a judge who is good at detecting the relevant normative properties (Ibid.).

I will not take a stand on whether these objections to dispositionalism, or any other objections, are decisive. But I do think it is misleading to contrast dispositionalism with neo-sentimentalism, where this latter view is taken to be distinctive in virtue of its focus on fittingness. Contrasting these views in this way rules out the possibility of defending a

dispositionalist account of fittingness.[23] But this is a possibility that should be left open, in particular because the most plausible versions of dispositionalism should also aspire to give analyses of fittingness. To see why, consider a dispositionalist analysis of the admirable in terms of what would elicit admiration from idealised subjects in idealised circumstances. Now, it is more or less a platitude that the admirable is what it is fitting to admire (given suitable qualifications to account for proportionality). The most plausible versions of dispositionalism will aim to respect this platitude. To do this, they could claim that the properties of *being such as to elicit admiration from idealised subjects in idealised circumstances* and of *being fitting to admire* are distinct, but necessarily co-extensive properties. But this would not be very attractive. For one thing, it would leave dispositionalists with the burden of explaining why these properties, despite being distinct, necessarily coincide with one another. A more attractive option would be to extend their dispositionalist analysis of the admirable to provide a dispositionalist analysis of fitting admiration as well.

Instead of seeing D'Arms and Jacobson's objections to dispositionalism as motivating a move away from dispositionalism to neo-sentimentalism (if successful), we do better to see them as motivating a move, within the neo-sentimentalist camp, away from dispositionalist versions of neo-sentimentalism to other versions – for example, versions on which fittingness judgements are given a non-cognitivist analysis (again, if these objections are successful). As indicated above, I will remain neutral on whether these, or any other, objections to dispositionalism are successful. Given this, my defence of MB is consistent with dispositionalist forms of neo-sentimentalism. Interpreted as a dispositionalist analysis, MB claims that deontic moral concepts and relations are analysable in terms of how certain,

[23] See Kauppinen (2014) for a defence of an ideal dispositionalist analysis of fittingness. Kauppinen also argues that Adam Smith (1982) defended an ideal dispositionalist analysis of fittingness in terms of the responses of what Smith called 'the impartial spectator'.

possibly idealised, individuals or groups would be disposed to respond with moral blame under certain, possibly idealised, circumstances.

1.2.2 Two Arguments for Neo-Sentimentalism

I begin with an argument due to D'Arms (2005) for accepting neo-sentimentalist analyses of certain normative concepts, before turning to an argument due to David Shoemaker (2017) and D'Arms and Jacobson (2017) for accepting neo-sentimentalist analyses of certain normative properties. Both arguments are best seen as schematic, in the sense that they outline ways of arguing for neo-sentimentalist analyses that need to be filled in by attending to the details of particular normative concepts and properties and the emotional responses that are associated with them. The first argument will be important to my defence of MB. The second will not, but since it has been influential in recent work on neo-sentimentalism it is worth introducing it to distinguish it from the approach I take here.

In 'Two Arguments for Sentimentalism', D'Arms develops what he calls 'the regulative role argument' for accepting neo-sentimentalist analyses of certain normative concepts.[24] The argument begins from an account of conceptual analysis that is very similar to the account I defended in 1.1. According to this account, it is not only a desideratum for conceptual analyses to respect central features of how we use and understand the concepts under analysis, but also that the analyses explain why the concepts in question are valuable ones that are worth keeping. D'Arms puts the point not primarily in terms of the value of concepts but rather in terms of their *functions*: it counts in favour of an analysis if it shows how the concept in question serves

[24] This is the first of two arguments he develops. The second is that neo-sentimentalists are especially well-placed to explain why certain evaluative disputes are univocal even though the evaluative concepts deployed in them are 'essentially contestable' (2005: 11-14). I will not discuss this argument in this thesis. Another argument sometimes given for neo-sentimentalism that I will not consider is that it is especially well-placed to explain 'judgment internalism': this is the claim, roughly, that (certain kinds of) normative judgments have an internal or necessary connection with motivation.

an important function that answers to our needs and interests (6). In D'Arms's view, neo-sentimentalists can take advantage of this desideratum on conceptual analyses by arguing that the normative concepts they seek to analyse serve the important function of *regulating our emotional responses*.

Normative concepts, as analysed by neo-sentimentalists, are able to serve this function because our emotional responses are significantly, albeit imperfectly, responsive to our judgments as to their fittingness (D'Arms 2005: 10-11). In some cases, this happens quickly: for example, if shocking new details about someone you previously admired come to light. In other cases, the process may take more time. For instance, someone might be afraid of their neighbour's large dog at first, but eventually overcome their fear as a result of judging that fear is not merited towards it (because of its sweet and gentle disposition). Moreover, judging that a kind of emotion that you do not typically feel in a given kind of circumstance would be fitting can, over time, dispose you to feel that kind of emotion in that kind of circumstance (5). Finally, our emotions are much less responsive to other kinds of normative judgments about them (for example, judgments about their prudential or moral value) than they are to fittingness judgments about them (11).

Emotions are not *perfectly* responsive to fittingness judgments. Emotions can be 'recalcitrant', as when someone's fear of flying persists despite them judging that such fear is not merited (D'Arms and Jacobson 2003). But the fact that our emotions are significantly, even though imperfectly, responsive to our fittingness judgments means that normative concepts, as analysed by neo-sentimentalists, are able to serve the function of regulating our emotional responses. Deploying these normative concepts will, over time and *ceteris paribus*, tend to bring about a loose harmony between our normative judgments and our emotional dispositions, such that we are, e.g., disposed to feel indignation towards, and only towards (more or less), that which we judge to be culpably morally wrong.

D'Arms argues that this function is important because we have good reasons for thinking about and discussing standards for what emotional responses to have (2005: 8-10). He focusses in particular on the ties between emotions and motivation (8-9), and the epistemic value of emotions (9-10). Emotions are typically motivating, and because it matters to us how we act, we have reasons for reflecting on standards concerning what to feel. Moreover, given that we often need to coordinate our actions with others, it is useful to have communicative devices for discussing, and ideally reaching agreement on, standards for what emotional responses to have (9). Emotions can also have epistemic value. One reason for this is that, because they can conflict with our normative judgments, our emotions can prompt critical reflection on those judgments (Ibid.). For example, someone who has been brought up in a racist society may be led to question their racist views by persistent, recalcitrant guilt-feelings concerning their treatment of racial minorities. Terms and concepts for assessing the fittingness of emotions facilitate such critical reflection.

D'Arms's regulative role argument is best seen as an argument schema that needs to be filled in by looking at particular normative concepts, and associated emotional responses, in detail. There are two reasons for this. First, the regulative role argument presupposes an account of conceptual analysis on which it is a desideratum that analyses of concepts are broadly in line with the concepts we in fact use. Showing that this desideratum is met will require considering the details of particular normative concepts and emotional responses. Second, showing that it is valuable to have concepts that allow us to regulate certain emotions by making fittingness judgments about them will surely require looking at the emotions in question in detail (even if reflection on general features of emotions plays some role). For example, if an emotion is associated with an extremely pernicious motivation that is almost always harmful, then it seems we might well conclude that we would be better off trying to expunge it as far as we can, rather than regulate it by making judgments about its fittingness.

In Chapter 3, I will aim to fill in these details for deontic moral concepts and moral blaming emotions by arguing that the account of deontic moral concepts provided by MB explains why these concepts serve extremely valuable social functions, chiefly because of the motivations associated with moral blaming emotions. The backdrop for these arguments will be the regulative role argument: deontic moral concepts are valuable insofar as they allow us to regulate moral blaming emotions by making judgments about their fittingness, and the value of such regulation, in turn, is explained in terms of certain features of moral blaming emotions – mainly, but not exclusively, the motivations associated with them.[25]

The regulative role argument outlines an approach to defending neo-sentimentalist analyses of normative *concepts*. But it is hard to see how it could support neo-sentimentalist analyses of the properties or relations these concepts pick out. Presumably, normative concepts could still be valuable in virtue of allowing us to regulate our emotional responses by fittingness judgments even if these concepts refer to the properties in virtue of which certain emotional responses are fitting, rather than to the property of being a fitting object of these emotions. The next argument I will examine is explicitly presented as an argument for neo-sentimentalist analyses of properties or relations (Shoemaker 2017; cf. also D'Arms and Jacobson 2017). Similarly to the regulative role argument, it is best understood as an argument schema that needs to be filled in by attending to the details of particular normative properties or relations and emotional responses.

The argument, in brief, is that it is not possible to give a plausible account of what unifies the instances of a given property or relation under analysis without appealing to the

[25] The account of the relation between fittingness judgments and our emotional responses on which the regulative role argument rests – according to which our emotions are significantly, but imperfectly, responsive to such judgments – is not uncontroversial. In particular, it stands in opposition to views on which our normative judgments generally are typically post-hoc rationalisations of prior emotional responses (cf. Haidt 2012 for a defence of this view primarily with respect to moral judgments). For extended criticism of such views, see, e.g., D'Arms and Jacobson 2023: 40-62).

fittingness of certain emotional responses (Shoemaker 2017: 487-493; cf. also D'Arms and Jacobson 2017: 254-255). Consider funniness, for example. Many different kinds of things can be funny. Clever wordplay, good impersonations (or bad ones), physical comedy. Defenders of neo-sentimentalist analyses of the property of being funny in terms of fitting amusement have argued that we cannot give a compelling rationale for what ties together the diverse instances of funniness without adverting to the fittingness of amusement (Shoemaker 2017: 487-493; D'Arms and Jacobson 2017: 254). Attempts to do so inevitably run into counter-examples. The funny cannot be reduced to the incongruous, since something can be funny without being incongruous and *vice versa* (Shoemaker 2017: 487; D'Arms and Jacobson 2017: 254). Nor can the funny be reduced to benign norm violations, since this proposal also faces numerous counter-examples (Shoemaker 2017: 487). We find the unity among instances of the funny only when we turn to the fact that they are all fitting objects of amusement. On the reasonable assumption that metaphysical analyses should reveal what unifies the instances of a given property or relation, it follows that we should accept a neo-sentimentalist analysis of the property of being funny.

Someone might try to argue for MB in a similar way. They could point to the considerable diversity in the instances of moral wrongness and argue that only MB reveals the unity among them. While this is a possible way of arguing for MB as a metaphysical analysis, it is not the argument I will pursue here. Such an argument would require taking on substantial commitments in first-order moral theory, since it would involve arguing that no first-order moral theory (or, at least, no first-order moral theory that posits a single basic moral principle) is correct. Although my argument for MB as a metaphysical analysis will appeal to its ability to provide unifying explanations, the explanations are at a different level: MB fits into a plausible general pattern of metaphysical analysis that explains what unifies different kinds of requirements across different normative domains. This is consistent with claiming that there is

a correct first-order moral theory that posits a single basic moral principle. I defend this argument in Chapter 2.

1.2.3 Two Problems for Neo-Sentimentalism

There are some general difficulties with neo-sentimentalist analyses that apply to MB in virtue of it being an analysis of this kind. One such difficulty is 'the wrong kind of reasons problem'. Another is 'the distance problem'. Let me introduce each of these problems and explain how they can be overcome. As we will see, the wrong kind of reasons problem arises only for neo-sentimentalists who analyse fittingness in terms of reasons. Since I remain largely neutral on how fittingness is best understood, this problem is orthogonal to my defence of MB. However, the distance problem is a more serious problem for my defence of MB, and accordingly I discuss it at greater length.

The wrong kind of reasons problem was originally introduced to the literature as a problem for 'buck-passing' accounts of value, which attempt to explain what it is for something to be valuable in terms of there being reasons to value it (Rabinowicz and Rønnow-Rasmussen 2004). The problem, in brief, is that there seem to be reasons for valuing things that are not valuable, and reasons for not valuing things that are valuable. For example, it seems that someone who is stuck in a boring, pointless job that has no value may nevertheless have reasons for valuing it if valuing it would make them happier. Buck-passers, it seems, owe us a non-circular account of what distinguishes reasons for/against valuing things that bear on whether they are/are not valuable ('reasons of the right kind') from reasons for/against valuing things that do not bear on this ('reasons of the wrong kind'). To supply such an account would be to solve the wrong kind of reasons problem.

Some buck-passers attempt to meet the wrong kind of reasons problem head-on by providing a non-circular account of what distinguishes reasons of the right kind from reasons of the wrong kind (Rabinowicz and Rønnow-Rasmussen 2004; Lang 2008; Schroeder 2010; Rowland 2019: 120-128). We need not go into the details of these attempts here. Other buck-passers attempt to dissolve the wrong kind of reasons problem by arguing that putative reasons of the wrong kind are not really reasons for/against valuing things, but rather reasons for different responses, such as *wanting* to value things or *trying* to value them (Skorupski 2010; Rowland 2019: 103-120).

Although the wrong kind of reasons problem has mostly been discussed in the context of buck-passing accounts of value, a closely analogous problem arises for versions of neo-sentimentalism that involve reasons-based analyses of fittingness. Consider, for example, a view according to which the admirable is analysed in terms of fitting admiration, and fitting admiration is analysed in terms of reasons to admire. It seems that there can be reasons for admiring things that would not be fitting to admire, and reasons for not admiring things that would be fitting to admire. The example given in 1.2.1, of an evil demon who threatens to torture you unless you admire something that is not admirable, seems to illustrate this (cf. Rabinowicz and Rønnow-Rasmussen 2004: 407-408). So it seems that neo-sentimentalists who understand fittingness in terms of reasons owe us a non-circular account of what distinguishes reasons for/against feeling emotions that bear on whether these emotions would be fitting (reasons of the right kind) from reasons for/against feeling emotions that do not bear on this (reasons of the wrong kind). To avoid circularity, such an account could not make essential reference either to fittingness or to the normative concepts or properties that are the target of the neo-sentimentalist analyses in question. Arguably, this problem is more than just analogous to the wrong kind of reasons problem that faces buck-passers: on the reasonable assumption that being valuable coincides with being fitting to value, many attempts to solve the wrong

kind of reasons problem for buck-passers can straightforwardly be adapted as proposals for solving the wrong kind of reasons problem for this variety of neo-sentimentalism (cf. McHugh and Way 2022: 81).

As I explained in 1.2.1, my defence of MB is neutral between many accounts of the fittingness relation, including reasons-based analyses, value-based analyses, and the view that fittingness is normatively primitive. Given this, the wrong kind of reasons problem is not a problem for MB as such, but only a problem for views that try to pair MB with reasons-based analyses of fittingness. I will remain neutral on whether this problem can successfully be solved or dissolved.

A second problem for neo-sentimentalist analyses is 'the distance problem'. The problem, in brief, is that it seems that various kinds of 'distance' between a subject and an object (e.g., temporal, spatial, personal, and modal) can make a difference to how it would be fitting for that subject to respond emotionally to that object, without making a difference to whether that object falls under the concepts/instantiates the properties with which neo-sentimentalists are concerned (Tappolet 2016: 105-110). For example, perhaps it can be fitting, over time, to feel less saddened by the loss of something valuable (for example, the death of a beloved pet), even though the loss does not become less sad over time (cf. Na'aman 2021). And perhaps how indignant it is fitting for you to feel towards someone for φ-ing can depend on your relation to those adversely affected by φ-ing, and also by your relation to the agent.[26] For example, perhaps the degree (or range of degrees) of indignation that would fittingly be

[26] Some might think that variation of this kind is ruled out by the nature of indignation. Indignation, it might be suggested, is by definition an impartial form of anger. While this is one way to use 'indignation', it is not how I will use it. To anticipate, in Chapter 2 I provide an account of indignation according to which it is a special form of anger that is distinguished by its motivational goal. The motivational goal of indignation is to get offenders to hold themselves accountable by feeling guilty (for the right reasons) about what they have done with respect to some other person(s) (that is, some other person(s) than the blaming agent) or impersonal value and make amends in the appropriate way. It is entirely possible that the fittingness of indignation, on this understanding of it, could be sensitive to facts about your relations to those adversely affected by the offender's φ-ing.

felt by the parent of a child who has been hurt by someone is much higher than the degree of indignation that would fittingly be felt by a stranger towards the same event. But many think that severity of moral wrongness is a non-relative matter, rather than relative to the positions of different agents.

There are three (and only three) possible approaches to addressing the distance problem. The first is to: (1) argue that fitting emotional responses and the concepts/properties with which neo-sentimentalists are concerned are impacted by distance in exactly the same ways. (Someone taking this line might argue that neither is impacted by distance at all). The second is to: (2) concede that fitting emotional responses and the concepts/properties with which neo-sentimentalists are concerned are impacted by distance in different ways, but attempt to defend more sophisticated neo-sentimentalist analyses that accommodate this point. The third is to: (3) take a mixed approach, by deploying (1) with respect to some pairs of fitting emotional responses and concepts/properties and (2) with respect to others. I will highlight some strengths of each of these approaches, and answer some objections, before ultimately arguing that the third, mixed strategy is most likely to succeed.

To pursue (1), we would need to show that, for any kind of distance (e.g., temporal, spatial, personal, or modal), *either* it affects both fitting emotional responses and the concepts/properties with which neo-sentimentalists are concerned (in the same way) *or* it affects neither of them. As an illustration of the first strategy, consider fitting sadness and the sad. Suppose, for the sake of argument, that the degree of sadness that is fitting in response to the death of a beloved pet can diminish over time. It may also be that the death of the pet can become less sad over time – for instance, when the owner gradually adjusts to the loss of the pet, by taking up new activities and perhaps getting a new pet. If this is right, and if diminishment in fitting sadness over time generally matches diminishment in how sad

something is over time, then an analysis of the sad in terms of fitting sadness is not threatened by temporal distance.

It is also open to a defender of (1) to argue that distance of a certain kind affects neither a given fitting emotion or its associated normative concept/property. Central to this strategy will be distinguishing fittingness from other kinds of normative assessment. For example, someone might concede that severity of moral wrongness is a non-relative matter, but try to accommodate the intuition that stronger indignation is in some sense 'appropriate' on the part of the parent by arguing that feeling such indignation is part of being a good parent, or perhaps that there are moral reasons to feel such indignation, even though it would not be fitting.[27] If this is right, then an analysis of moral wrongness in terms of fitting moral blame is not threatened by distance of this kind ('personal' distance), because such distance impacts upon neither moral wrongness nor fitting moral blame.

Perhaps these examples can be dealt with in these ways (although I will voice some misgivings about this way of dealing with the indignation example in a moment). But if (1) is to be plausible as a *general* response to the distance problem, then we need some *general* reasons for thinking that distance always impacts fitting emotional responses and the concepts/properties with which neo-sentimentalists are concerned in the same ways. One way of arguing for this, and possibly the only way, is to argue that the equivalences between fitting emotions and the normative concepts/properties that neo-sentimentalists seek to analyse provide our principal way of latching on to fittingness as a distinctive way of normatively assessing emotions. For example, Chris Howard writes: 'the equivalences serve to distinguish fittingness from other normative notions, grounding our grasp of the relation's extension, and giving our judgments about fittingness—rather than oughts or reasons, for instance—their

[27] This is similar to the response that Olson (2009: 373-377) gives to an analogous problem facing fitting attitude analyses of value.

distinctive content' (forthcoming: Sec. 2, fn. 5). And also: 'I lose any sense of what being fitting to regret, or what being regrettable, could be, if we allow that something could be fitting to regret without being regrettable (or vice versa)' (Sec. 4, Para. 14). If this is right, then distance must impact fitting emotions and the concepts/properties with which neo-sentimentalists are concerned in exactly the same ways.

However, this way of responding to the distance problem faces two problems. First, it might be wondered whether the postulated equivalences between, e.g., fitting admiration, regret, and shame, and the admirable, regrettable, and shameful, are even correct. For example, claims like 'It is fitting to be ashamed of X if and only if X is shameful', without further qualification, seem straightforwardly false (D'Arms and Jacobson 2023: 71-77; Achs and Na'aman forthcoming). One problem concerns proportionality: strong shame over something that is only mildly shameful would not be fitting. Another issue concerns the position of the agent who has the response: it would not be fitting for me to feel ashamed of a shameful trait belonging to someone who is a perfect stranger to me. Other examples of similar unqualified biconditionals face further problems. To give a non-emotional (but still conative) example, consider 'It is fitting to prefer X to Y if and only if X is preferable to Y'. Preferability often seems to be an agent-relative matter: something can be preferable for a particular agent, or group of agents, without being preferable for everyone. For example, it might be preferable for Alex that his partner wins a race, but this would not be preferable for everyone – for example, it would not be preferable for the other contestants.[28]

Examples like this show, not that we should reject the postulated equivalences, but rather that they need to be suitably qualified. In particular, we need to include a clause

[28] For further examples of problems with unqualified biconditionals like, 'It is fitting to be ashamed of X if and only if X is shameful', and, 'It is fitting to prefer X to Y if and only if X is preferable to Y', see Achs and Na'aman (forthcoming).

concerning proportionality and a clause concerning the agents for whom the responses in question would be fitting, as follows:

> *Qualified Fit/R-Able Biconditionals*: It is fitting for an agent, A, to have a conative response, R, of strength, S, to an object, O, if and only if O is R-able for A, and the degree to which O is R-able for A is proportional to the degree of S.[29]

When qualified in these ways, the biconditionals stated above are no longer subject to the counter-examples we considered. Strong shame would not be fitting to something mildly shameful, because it would fail to meet the proportionality condition on conative fit. Moreover, it would not be fitting for me to feel ashamed of a shameful trait belonging to someone who is a perfect stranger to me, given that their shameful trait is not shameful for me.[30] And finally, a suitably qualified biconditional concerning the link between fitting preference and preferability can register the point that fitting preference and preferability are often agent-relative matters.[31] Note, moreover, that qualifying the biconditionals in these ways still allows us to claim that the equivalences between fitting emotions and properties like shamefulness, admirableness, and amusingness provide our principal way of latching on to fittingness as a distinctive way of normatively assessing emotions. This is because the qualifications introduced to the right-hand-side of the biconditionals, concerning the proportionality and possible agent-relativity of R-

[29] It may be that further qualifications are needed concerning time. See Achs and Na'aman (forthcoming) for discussion. (It should be noted, however, that Achs and Na'aman argue that even biconditionals that are qualified in these ways are still open to counter-examples.)

[30] D'Arms and Jacobson suggest that claiming that, e.g., cowardice is shameful for cowards, but not for people who are not cowards or saliently related to cowards, is 'misleading' (2023: 73). However, even if this claim is misleading or odd, this does not mean it is not true. Moreover, there are contexts in which it would be natural to utter these kinds of statements. For example, imagine that someone feels ashamed of a shameful trait possessed by someone who is not especially close to them. It seems that we might well say, 'so-and-so's trait is shameful/embarrassing for them, but not for you'.

[31] Achs and Na'aman (forthcoming) raise some further problems for biconditionals of this kind. While it would take me too far afield to discuss these issues, I believe that their objections are unsuccessful. One central kind of counter-example they give concerns the duration of responses. They suggest, for example, that it may not be fitting for someone to continue to be amused by something when they have already been amused by it for a long time, even if it is genuinely amusing. But cases of this kind seem to be covered by the proportionality condition in the qualified bi-conditionals. Being amused by a joke that is only moderately amusing for too long would be unfitting because it would be an overreaction and hence disproportionate.

able properties, are ones on which we have an independent grasp. We have an independent grasp on the idea that one trait may be more shameful than another, and that something can be preferable for one person but not for someone else.

However, there is a second problem with responding to the distance problem by relying on the claim that the equivalences between fitting emotions and properties like shamefulness, admirableness, and amusingness provide our principal way of latching on to fittingness as a distinctive way of normatively assessing emotions. The problem is that this immediately raises a concern about circularity. The concern is that neo-sentimentalism is circular insofar as it attempts to analyse various normative concepts/properties in terms of fitting emotions, but then explains the relevant notion of fittingness in terms of the very normative concepts and properties that it seeks to analyse. Circularity of this kind might not be fatal to the neo-sentimentalist – indeed, as we saw in 1.2.1, some neo-sentimentalists explicitly defend circular analyses.[32] But this would leave us with the difficult task of explaining how circular analyses can nonetheless be informative. Other things equal, it would be better to respond to the distance problem without incurring circularity.

Perhaps we can grant that the equivalences between fitting emotions and the concepts/properties that neo-sentimentalists seek to analyse – or, at least, *some* of these equivalences – provide our principal way of latching on to fittingness as a distinctive way of normatively assessing emotions without rendering neo-sentimentalist analyses circular. Two points are especially worth making. First, as we saw in 1.2.1, many terms for emotions have associated normative terms formed by suffixation, such as 'admiration'/'admirable', 'shame'/'shameful', 'amusement'/'amusing', and 'regret'/'regrettable'. It is open to neo-

[32] However, the circularity in question is usually held to be that the relevant *emotional responses* cannot be understood independently of the normative concepts/properties under analysis, rather than that the relevant notion of *fittingness* cannot be so understood. (Cf. McDowell 1985; Wiggins 1987).

sentimentalists to argue that the connections between *these* normative terms (or, alternatively, the concepts they express or the properties to which they refer) and fitting emotional responses provide our primary way of latching on to fittingness, but that, once we have latched on in this way, we can confidently go on to apply the notion of fittingness to other responses that lack such terms. This suggests that, even if these normative terms serve as our entry point to the notion of fittingness, we are later able to apply it independently of them. An illustration of this is 'resentment', which lacks an associated normative term formed by suffixation. It seems that we can confidently identify a kind of normative success that resentment possesses insofar as it stands in the same kind of relation to its object as, e.g., admiration stands in to its object when its object is admirable. This suggests that terms like 'shameful', 'admirable', and so on allow us to home in on a relation that we can later identify independently of them.

A second point that is worth making is that neo-sentimentalists can still claim that what being admirable, amusing, contemptible, and so on *have in common* is their connection with fitting emotional responses, even if the connection between these properties and fitting emotions is what initially orients us to the relevant notion of fittingness. Putting this point together with the point made in the last paragraph, defenders of the response to the distance problem under consideration can respond to the circularity objection at issue as follows. The objection, recall, is that neo-sentimentalism is circular insofar as it attempts to analyse various normative concepts/properties in terms of fitting emotions, but then explains the relevant notion of fittingness in terms of the very normative concepts and properties that it seeks to analyse. Neo-sentimentalists can respond by denying that they are *explaining* the relevant notion of fittingness in terms of these concepts and properties. Rather, these normative concepts and properties – or, at least, some of them – provide us with our initial handle on the relevant notion of fittingness, but the fittingness relation is one on which we subsequently have an independent grip, as evidenced by our ability to apply it to new cases. Moreover, insofar as we can appeal

to fittingness to explain what these different normative concepts and properties have in common, neo-sentimentalist analyses of them are appealing.[33]

It might be wondered, however, how this response to the distance problem will help us with neo-sentimentalist analyses of normative concepts and properties/relations beyond the admirable, the contemptible, the amusing, the regrettable, and so on. In particular, how will it help us with MB? This is where response (1) to the distance problem (i.e., to argue that fitting emotional responses and the concepts/properties with which neo-sentimentalists are concerned are impacted by distance in exactly the same ways) falls short. The problem is that the tight link between, e.g., fitting fear, shame, and preference, and the fearsome, shameful, and preferable, seem to show that fittingness typically depends on factors that drive a wedge between moral wrongness and the fittingness of moral blaming emotions. Earlier, we asked whether the degree of indignation that would fittingly be felt by the parent of a child who has been hurt by someone is higher than the degree of indignation that would fittingly be felt by a stranger towards the same event. An affirmative answer to this question would threaten MB, insofar as many think that severity of moral wrongness is a non-relative matter, rather than relative to the positions of different agents.

Now, it seems that fearsomeness, preferability, and perhaps even shamefulness are often sensitive to considerations to do with partiality. As we saw earlier, it might be preferable for Alex that his partner wins a race, but this would not be preferable for everyone. Similarly, the prospect of Nancy dumping her boyfriend might be fearsome for Paul, if Paul is Nancy's boyfriend, but not fearsome for a stranger. And finally, although this is a more difficult case,

[33] It is worth pointing out that the claim relied on here – *viz.*, that some of the normative concepts/properties that neo-sentimentalists seek to analyse provide us with our initial grip on the fittingness relation – is different from some of the claims that Howard (forthcoming) makes in the passages quoted above. In particular, it is different from the claim that the equivalences between fitting emotional responses and certain normative concepts/properties give 'our judgments about fittingness… …their distinctive content' (forthcoming: Sec. 2, fn. 5). I am not sure whether this claim can be squared with the possibility of defending non-circular neo-sentimentalist analyses.

perhaps it might be shameful for Andrew if his country commits awful war crimes, but this would not be shameful for someone who is not a citizen of Andrew's country or saliently connected to it. Insofar as these examples suggest that fittingness is typically sensitive to considerations to do with partiality, this suggests that it *would* be fitting for the parent of a child who has been hurt by someone to feel much stronger indignation than would be fitting from a stranger – exactly the wrong result for MB. So, even if the first way of responding to the distance problem (i.e., to argue that fitting emotional responses and the concepts/properties with which neo-sentimentalists are concerned are impacted by distance in exactly the same ways) works for neo-sentimentalist analyses of the amusing, shameful, admirable, and the like, it seems that it will not help us with neo-sentimentalist analyses of other kinds of normative concepts and properties/relations, such as deontic moral concepts and relations.

Let me turn then to the second way of responding to the distance problem. This is to concede that fitting emotional responses and the concepts/properties with which neo-sentimentalists are concerned are impacted by distance in different ways, but attempt to defend more sophisticated neo-sentimentalist analyses that accommodate this point.

A good illustration of this approach is Allan Gibbard's (1990) defence of a neo-sentimentalist conceptual analysis of MORALLY WRONG partly in terms of MORALLY BLAMEWORTHY, where MORALLY BLAMEWORTHY, in turn, is understood in terms of FITTING ANGER AND GUILT. Gibbard claims that spatial distance sometimes makes a difference to our judgments about the fittingness of anger (or, in his preferred terminology, our judgments as to whether anger is 'warranted', 'apt', or 'makes sense') (126). He gives the example of theft of a camel: it might be fitting for you to be very angry if someone steals *your* camel, but it would be strange for you to be as angry if a far-away stranger's camel was stolen (Ibid.). But how morally blameworthy the thief is, and how seriously morally wrong their action is, seem not to vary depending on your personal involvement in the situation.

Gibbard suggests that we can nonetheless analyse MORALLY WRONG in terms of FITTING ANGER AND GUILT by distinguishing the question of what is fitting to feel *given full engagement with a certain standpoint*, and the question of whether *we have reasons to engage fully with that standpoint* (cf. 127). His proposal, roughly, is that we can analyse the concept of moral wrongness in terms of whether anger would be fitting given full engagement with an impartial standpoint, and whether guilt would be fitting given full engagement 'in an aspect of one's own place in the world' (127).[34] Returning to the camel example, you might have good reasons for fully engaging with a personal (or partial) standpoint when your camel is stolen, but not when a far-away stranger's camel is stolen. It is fitting for you to be very angry when your camel is stolen given full engagement with *your* personal standpoint, but it would be strange for you to be as angry about the theft of the stranger's camel because you do not have good reasons to engage fully with *their* personal standpoint. In Gibbard's view, when we judge whether an action is morally wrong, and, if so, how seriously morally wrong it is, we adopt a common, impartial standpoint, and consider whether it would be fitting to be angry given full engagement with that standpoint.[35]

Gibbard's analysis exemplifies response (2) to the distance problem distinguished above (i.e., to concede that fitting emotional responses and the concepts/properties with which neo-sentimentalists are concerned are impacted by distance in different ways, but attempt to defend more sophisticated neo-sentimentalist analyses that accommodate this point). One general strength of this strategy is that it avoids the risk of circularity that besets response (1). However, the cost of this is working with a more obscure notion of fittingness. Moreover, insofar as we have no antecedent reason to expect that different neo-sentimentalist analyses

[34] This is rough, because Gibbard's considered statement of the analysis also includes a clause concerning responsibility. See (1990: 40-45).

[35] The claim that moral judgments require us to adopt an impartial standpoint has a long history in the sentimentalist tradition. Early examples are David Hume's 'common point of view' and Adam Smith's 'impartial spectator'. For discussion of Hume's 'common point of view', see, e.g., Cohon (1997) and Driver (2012: 282-285), and for discussion of Smith's 'impartial spectator', see, e.g., Sayre-McCord (2010).

will require the same adjustments, response (2) might commit us to giving piecemeal responses to the distance problem for different neo-sentimentalist analyses.

The best strategy may well be to combine (1) and (2), by deploying these approaches with respect to different normative concepts and properties. In particular, we might claim, with response (1), that the normative concepts and properties that are most immediately and obviously connected with fitting emotional responses (amusing, shameful, contemptible, and the like), are such that distance impacts these normative concepts and properties and their associated fitting emotions in exactly the same ways. Insofar as we have a clear, intuitive grasp on these normative concepts and properties, this provides us with an initial orientation to the notion of fittingness as a distinctive form of normative assessment. But we might then agree with response (2) that other normative concepts and properties that are less immediately and obviously connected with emotional responses are such that neo-sentimentalist analyses of them need to be complicated in light of the distance problem. Deontic moral concepts and relations are one example here, but other examples may include virtue, vice, and well-being. Indeed, we have already seen good reasons for thinking that, because of the issue of partiality, how morally blameworthy someone is, and how seriously morally wrong their action is, can come apart from how indignant it would be fitting for certain agents to feel (even when the offender acts without a moral excuse). It therefore seems necessary for defenders of MB to incorporate something like Gibbard's appeal to fittingness *from an impartial standpoint* into the account of moral blameworthiness that figures in their analysis.

The relevant notion of an impartial standpoint is simply the standpoint of an agent who was not personally affected by the object of assessment – neither they, nor those with whom they stand in special relationships, were disadvantaged (or advantaged) by the offender's φ-ing. It is worth emphasising that appealing to an impartial standpoint in this way is consistent with thinking that there are moral requirements to be partial towards those with whom we stand

in special relationships. Indeed, it is consistent with thinking that we have *underivative* moral requirements of this kind – moral requirements that do not derive from fundamentally impartial requirements. This is because it is possible to hold that indignation from an impartial standpoint would be fitting towards unexcused violations of standards enjoining agents to be partial towards those with whom they stand in special relationships. Moreover, it is possible to hold that the fittingness of such indignation does not derive from some fundamentally impartial standard that is violated in such cases.

1.3 Chapter Summaries

Chapters 2 and 3 lay out my main positive arguments for MB. Chapter 2 explains the various elements of MB, argues that MB satisfies the descriptive desideratum on conceptual analyses, and argues that it provides a plausible analysis of deontic moral relations as well as concepts. The platitudes I focus on concern the link between moral wrongness and making amends, the claim that moral wrongs can be more or less serious, and the link between being a morally responsible agent and being subject to moral requirements. I argue that MB gains plausibility as a metaphysical analysis in virtue of conforming to an attractive general pattern of analysis of different kinds of requirements, permissibility, and wrongness across different normative domains (for ease of presentation, I focus mainly on requirements). MB thus forms part of an attractive explanation of what unifies different kinds of requirements.

Chapter 3 argues that MB satisfies the prescriptive desideratum on conceptual analyses. I argue that the account of deontic moral concepts provided by MB explains how deontic moral concepts serve two valuable social functions: what I call 'the Deliberative Reliability Function' and 'the Respect Function'. By 'the Deliberative Reliability Function', I mean the function of protecting important interests through encouraging reliable patterns of deliberation and

motivation. Examples of these interests include the interests we have in not being killed, hurt, or lied to (and the further interests we have in being able to rely on not being treated in these ways). By 'the Respect Function', I mean the function of providing social recognition of the dignity of persons and their worthiness of respect.

Chapter 4 discusses objections to the extensional adequacy of MB. According to these objections, MB cannot provide a plausible conceptual and metaphysical analysis of moral wrongness, because the left-hand side of MB does not necessarily coincide with its right-hand side. I address various extensional challenges to MB and consider what changes, if any, need to be made to MB to meet them. I argue that, with the exception of one extensional challenge, no changes need to be made to MB. Moreover, defenders of MB can meet these challenges while remaining largely neutral on the first-order debates at issue. The extensional objections I discuss arise from the Objectivism/Subjectivism debate, the debate surrounding the possibility of suberogation, the debate surrounding the relevance of motivating reasons to moral wrongness, and the consequentialist tradition. I argue that, with the exception of challenges arising from the Objectivism/Subjectivism debate, no changes need to be made to MB to meet them successfully.

Chapter 5 focusses on circularity objections to MB. The most straightforward circularity objection to MB is generated by pairing the following account of what it is for emotions to be fitting with the following claim about the nature of indignation, the central moral blaming emotion:

> *Emotional Fittingness as Accurate Representation*: For an emotion to be fitting is for it to be an accurate representation of its object.
>
> *Deontic Moral Content*: Indignation involves a judgment or other attitude including the content: *it is morally wrong for the agent to φ*.

If we plug these claims back into MB, we get: for it to be morally wrong for an agent to φ is for φ-ing to violate standards such that, if the agent violated those standards without a moral excuse, the pair <agent, φ-ing> would be accurately represented as instantiating the moral wrongness relation (among other relations). This is clearly circular.

I argue that we have good reasons for rejecting both Emotional Fittingness as Accurate Representation and Deontic Moral Content. Against Emotional Fittingness as Accurate Representation, I argue that the fittingness of an emotion as a whole is a function of the fittingness of both its representational and motivational components. And against Deontic Moral Content, I argue, first, that indignation may have no normative content whatsoever, and, second, that even if indignation has (non-moral) normative content, it is not plausibly understood as having specifically deontic moral content.

Chapter 6 discusses further non-extensional objections to MB besides circularity objections. I focus on three objections. The first is that those whom I call guilt and/or resentment and indignation 'rejecters' – people who claim that guilt and/or resentment and indignation would rarely or never be fitting – pose problems for MB, insofar as rejecters might seem able to make deontic moral judgments without contradicting themselves or suffering from conceptual confusion. The second objection concerns a certain kind of 'Moral Twin Earth' thought experiment. And the third objection is that MB cannot adequately account for the moral praiseworthiness of actions performed from the motive of moral obligation. I argue that all of these objections are unsuccessful. Moreover, seeing why the third objection is unsuccessful reveals a further strength of MB: MB fits nicely with an independently plausible view concerning the conditions of moral praiseworthiness.

Finally, Chapter 7 considers the upshots of MB for the connection between moral wrongness and normative reasons. Some philosophers have argued that MB supports the

Overridingness Thesis, according to which we always have decisive normative reasons not to act morally wrongly (Gibbard 1990: 299-300; Darwall 2006a: 97-99, 2006b: 292; Skorupski 2010: 295- 301; Portmore 2011: 38-51, 2021: 51-62). I argue that this argument fails. I go on to argue that MB supports only a weaker thesis I call 'the Normativity Thesis'. According to the Normativity Thesis, if it is morally wrong for an agent to φ, then that agent has significant normative reasons not to φ. However, whether these reasons are decisive or even sufficient is left open by this thesis.

Chapter 2

Arguments for Moral Wrongness as Moral Blameworthiness: Part 1

This is the first of two chapters laying out my main positive arguments for MB. This chapter explains the various elements of MB, argues that MB satisfies the descriptive desideratum on conceptual analyses explained and defended in the Introduction, and argues that it provides a plausible analysis of deontic moral *relations* as well as *concepts*. As we saw in the Introduction, MB holds:

> *Moral Wrongness as Moral Blameworthiness (MB)*: It is morally wrong for an agent to φ iff (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, where 'φ' stands for an object of deontic moral assessment.

There are various elements of this analysis that call out for explanation. What are standards? What is it for an agent to be morally blameworthy for violating standards? And what is a moral excuse? Section 2.1 answers these questions.

The descriptive desideratum on conceptual analyses is that they must be broadly in line with the concepts we in fact use. More precisely, successful conceptual analyses respect many of the platitudes surrounding the concepts under analysis, where 'platitudes' is a term of art for statements of the judgmental and inferential dispositions that are typically had by those who possess the relevant concepts. I argue that MB meets this descriptive desideratum in 2.2. I also provide a fuller characterisation of platitudes by explaining the notion of concept possession I have in mind, indicating some of the characteristic marks of platitudes, and explaining what it means for a conceptual analysis to 'respect' a platitude. The platitudes I discuss concern the link between moral wrongness and making amends, the claim that moral wrongs can be more

or less serious, and the link between being a morally responsible agent and being subject to moral requirements.

I defend MB not only as an analysis of deontic moral concepts, but also as a metaphysical analysis of deontic moral relations. Section 2.3 gives my argument for MB as a metaphysical analysis. I argue that MB gains plausibility as a metaphysical analysis in virtue of conforming to an attractive general pattern of analysis of different kinds of requirements, permissibility, and wrongness across different normative domains (for ease of presentation, I focus mainly on requirements). MB thus forms part of an attractive explanation of what unifies different kinds of requirements.

2.1 Clarifications

This section clarifies three aspects of MB. It takes up three questions. What are standards? What is moral blameworthiness? And what are moral excuses?

2.1.1 What are Standards?

A standard is a criterion by which one can, in principle, judge or appraise various things on the basis of meeting or failing to meet the condition set by the criterion (cf. Copp 1995: 19). Standards can pertain to various kinds of things, such as actions, character traits, and beliefs. Moreover, standards can pertain to various kinds of judgment or appraisal. For example, an agent may be open to appraisal as morally blameworthy if they fail to meet the condition set by a standard without a moral excuse. I use ‘φ-ing violates standards’ to mean that φ-ing is open, in principle, to negative judgment or appraisal on the basis of failing to meet the conditions set by the standards. Standards, as I understand them here, need not be general

principles or rules; they may be particularistic. Given this understanding of standards, MB is compatible with both the view that moral principles play important theoretical and/or practical roles in morality and the view that they do not. That is, MB is neutral with respect to the moral particularism/moral generalism debate.[1]

2.1.2 What is Moral Blameworthiness?

Blame has received a great deal of attention in recent years.[2] Like many, I take this attention to have revealed that blame is a highly diverse phenomenon.[3] There are a great many critical reactions that can reasonably be called blame. These reactions include guilt, resentment, and indignation, which have been a central focus in discussions of blame since P. F. Strawson's influential paper "Freedom and Resentment" (1962/2003). Blame also includes many other emotional and non-emotional reactions. For example, shame, contempt, moral disgust, regret, and the kinds of relationship modifications emphasised by T. M. Scanlon (2008) can all reasonably be considered forms of blame.

Defenders of MB and related views have tended to focus on guilt, resentment, and indignation (Gibbard 1990; Darwall 2006; Hooker 2017; Kauppinen 2017). This focus is well-motivated. As we will see shortly, guilt, resentment, and indignation are all closely linked with making amends. Intuitively, there is a close relation between *it being morally wrong for an agent to φ* and *there being a reason for that agent to make amends for φ-ing without a moral*

[1] This is a rough-and-ready characterisation of the moral particularism/moral generalism debate. For a more careful characterisation, see McKeever and Ridge (2006: 3-24).
[2] See Coates and Tognazzini (2013) for a collection of perspectives on blame. Some important recent discussions of blame include Wallace (1994); Sher (2005); and Scanlon (2008). Throughout, I use 'blame' to refer to both self-directed and other-directed forms of blame. Guilt is a kind of self-blame.
[3] 'Diverse' in the sense that various kinds of emotional and non-emotional critical reactions can constitute blame (compare, e.g., Shoemaker (2015); Fricker (2016); Nussbaum (2016: 256-260); and Mason (2019)).

excuse.[4] Indeed, I will argue in 2.2 that this is one of the platitudes surrounding MORALLY WRONG. In contrast, other forms of blame are *prima facie* less closely related to moral wrongdoing. For example, shame, contempt, and moral disgust all seem more closely tied to negative evaluative appraisals of relatively enduring features of agents, such as character traits, than deontic appraisal of actions as morally wrong.[5] Henceforth, by 'moral blame' I mean guilt, resentment, and indignation.

It will be important in what follows to have an account of the motivational goals of these emotions. By the 'motivational goals' of emotions, I mean the satisfaction conditions of their motivational components. For instance, just as a desire to eat an apple is satisfied by eating an apple, the motivational component of fear of a snake is satisfied by becoming safe from the snake. Insofar as fear involves a motivational component, the motivational goal of fear is to become safe from its object (Scarantino 2014: 169). (I say more about the kinds of motivational states that emotions characteristically involve in Chapter 5).

Focussing on the central case of guilt for actions or omissions, the motivational goal of guilt is to make amends for actions or omissions through such things as confessions, apologies, and offers of compensation (Tangney and Dearing 2002: 19; De Hooge 2019). Now, it may be that 'guilt' is sometimes used to refer to emotions that do not have this motivational goal. Perhaps 'buyer's guilt' is an example of this. I will bracket off such uses of 'guilt'. My claim that the motivational goal of guilt is to make amends, then, is partly stipulative: I am using 'guilt' to pick out an emotion that has this motivational goal. My uses of 'resentment' and 'indignation' are partly stipulative as well. I use 'resentment' and 'indignation' as terms of art

[4] Close, but not exceptionless: it is not the case that it is morally wrong for an agent to φ if and only if it is appropriate for that agent to make amends for φ-ing. As we will see shortly, it can be appropriate for an agent to make amends for φ-ing without it being morally wrong for that agent to φ, because φ-ing is a personal matter and warrants guilt and resentment but not third-personal indignation.

[5] For this point with respect to shame, see, e.g., Williams (1993: 92-93) and Deonna et al. (2011: 70-122); with respect to contempt, Mason (2003: 246-250) and Bell (2013: 37-48); and with respect to moral disgust, Giner-Sorolla et al. (2018).

for forms of anger that are distinguished from other forms of anger by their motivational goals. The motivational goal of resentment and indignation is to make offenders hold themselves accountable by feeling guilty (for the right reasons) and making amends for their conduct. Dill and Darwall (2014: 46-54) review extensive empirical evidence for the claim that these emotions are psychologically real and the typical driving force behind condemnatory behaviour. Resentment and indignation are distinguished by their motivational goals from other kinds of anger, such as vengefulness. I remain neutral as to how anger in general is to be understood and what unites its various forms. It may be that there is a common element to the motivational goals of the varieties of anger. Perhaps they all aim at confrontation in some broad sense.[6] On the other hand, it may be that the varieties of anger are more loosely connected by overlapping similarities or family resemblances.

I use 'resentment' and 'indignation' to distinguish personal from vicarious responses (cf. Strawson 1962/2003). This distinction can be elaborated in terms of a difference in the motivational goals of these emotions. Resentment aims at making the offender feel guilty (for the right reasons) about what they have done to *you* and make amends with *you*. Indignation, on the other hand, aims at making the offender feel guilty (for the right reasons) about what they have done with respect to some other person(s) or impersonal value and make amends in the appropriate way.

Among these three emotions, indignation is especially important to understanding the concept of moral wrongness. It seems plausible that there are actions for which guilt and resentment are fitting, but not indignation, and that such actions are morally permissible. A possible illustration of this is violation of standards concerning good friendship. It may be fitting for Bianca to resent Carl for failing to keep in touch and for Carl to feel guilty about this.

[6] See Shoemaker (2015: 87-117) for the claim that the motivational goal of (agent-directed) anger in general is confrontation.

But indignation seems out of place, as does talk of moral wrongness. A second reason why indignation is especially important to understanding the concept of moral wrongness is that an analysis of the concept of moral wrongness needs to leave room for the possibility of undirected moral wrongs. A moral wrong is directed when it wrongs someone; it is undirected when it is morally wrong but does not wrong anyone (Owens (2012: 44-67); Darwall (2013a: 20-39); Wallace (2019a: 5-11); Jonker (2020: 3-16). Since resentment is a personal reaction, to claim that an action is morally wrong only if resentment is fitting would imply that all moral wrongs are directed. This means that defenders of MB had better allow for the possibility of morally wrong actions for which indignation is fitting, but not resentment, because there is no-one who has been wronged.[7]

So much for moral blame. How should defenders of MB understand moral blame*worthiness*? There are many characterisations of moral blameworthiness in the literature. Some characterise moral blameworthiness in terms of whether agents would be *fitting targets* of moral blame (e.g., Graham 2014; Rosen 2015b; Shoemaker 2017). Others understand moral blameworthiness in terms of whether moral blame would be *deserved* (e.g., Pereboom 2014). Another view is that moral blameworthiness for φ-ing concerns whether φ-ing reveals attitudes that *impair the relationships* that others can have with you (Scanlon 2008). A different view is that moral blameworthiness is a matter of whether moral blame would be *morally fair* (Wallace 1994). And finally, there is a well-known tradition of understanding moral blameworthiness in (direct or indirect) *consequentialist* terms (e.g., Smart 1961; Vargas 2013).

[7] Although Gibbard often frames his analysis in terms of 'resentment', he is clear that what he ultimately has in mind is impartial anger: norms of moral right and wrong, he claims, concern whether it makes sense to feel anger from a perspective of 'full, impartial engagement' and guilt from a standpoint of 'full engagement in an aspect of one's own place in the world' (1990: 127). Darwall acknowledges that resentment, understood as a personal reaction, can be fitting in response to actions that are not morally wrong (2006a: 80-82). As we saw in the Introduction, because of complications arising from the 'distance problem' defenders of MB should appeal not simply to the fittingness of indignation but to its fittingness from an impartial standpoint.

Although some of these views – in particular, direct consequentialist views – seem highly revisionary with respect to our ordinary understanding of moral blameworthiness, many of these views can reasonably be claimed to capture what intuitively counts as a notion of moral blameworthiness. Rather than seeing these different views as necessarily in competition with one another, we might do better to see them as articulating different notions, or kinds, of moral blameworthiness, and ask about the different ways in which these notions are important to moral thought. Taking this approach, we can now ask which of these notions is best suited to figuring in MB.[8]

Understanding moral blameworthiness in terms of whether agents would be fitting targets of moral blame seems promising. As we saw in the Introduction, the fittingness relation is the relation that holds between an object, a subject, and a response when the object merits – or, equivalently, is worthy of – that response from that subject. For example, admiration is fitting when it is felt towards something (proportionally) admirable, and fear is fitting when it is felt towards something (proportionally) fearsome. It is generally held that the fittingness of a response is typically independent of the value of the consequences of that response (D'Arms and Jacobson 2000b; Rabinowicz and Rønnow-Rasmussen 2004). This is significant, since consequentialist accounts of moral blameworthiness (at least, direct consequentialist accounts) would seem a poor match for MB, insofar as moral wrongness does not always line up in the right way with optimific moral blame.[9] At the same time, fittingness is not a specifically moral

[8] This approach is controversial. Some philosophers may interpret the views outlined in the paragraph above as rival views concerning how *the* moral blameworthiness relation is to be understood, rather than as articulations of different notions, or kinds, of moral blameworthiness (cf. Graham 2014: 388-393). It would take me too far afield to consider the merits of these different methodological approaches. However, even if this rival methodological approach is the best one, and even if the account of moral blameworthiness in terms of whether agents would be fitting targets of moral blame is rejected, MB might be re-interpreted and defended as an analysis of moral wrongness directly in terms of fittingness of moral blame (which, under the current supposition, would be distinct from moral blameworthiness).

[9] However, see Miller (2014) for an argument that indirect consequentialist accounts of moral blameworthiness are compatible with MB. I discuss the relation of MB to the consequentialist tradition at length in Chapter 4 and endorse Miller's claim there.

relation. In this respect, it contrasts with moral fairness, and perhaps also desert, insofar as desert of blame is sometimes thought to be closely connected with justice (see, e.g., Portmore 2022: 53). Understanding moral blameworthiness in terms of the fittingness of moral blame, then, allows defenders of MB to avoid the extensional objections that would beset their view given a (direct) consequentialist account of moral blameworthiness, while deploying a normative relation that is sufficiently independent of moral wrongness to promise an illuminating analysis of it.

Before I move on, it is worth addressing an objection to the claim that accounts of moral blameworthiness in terms of whether agents would be fitting targets of moral blame succeed in capturing even one notion of moral blameworthiness among others. The objection is that any adequate account of moral blameworthiness needs to distinguish whether agents are morally blameworthy for φ-ing from whether a given blamer has standing to blame these agents morally for φ-ing (Smith 2007). A standard example is hypocrisy: there seems to be something objectionable about hypocritical moral blame, but whether it would be hypocritical for a blamer to blame an agent morally, and whether that agent is morally blameworthy, are two different questions (Smith 2007: 479-480; Scanlon 2008: 176-177). What is objectionable about hypocritical moral blame seems rather to be that the blamer in some sense lacks standing to blame morally (Ibid.). Fittingness, as we have seen, is a three-place relation between a subject, an object, and a response, so accounts of moral blameworthiness in terms of whether agents would be fitting targets of moral blame threaten to collapse the distinction between moral blameworthiness and standing to blame morally.

There are two ways in which we might deal with this objection, depending on the precise source(s) of the objectionableness of morally blaming someone when you lack standing. The objection to morally blaming someone if you lack standing might at bottom be a moral objection: for example, perhaps hypocritical blamers are morally bad, or hypocritical blame is

morally wrong (cf. Wallace 2010). Now, we saw in the Introduction that there are reasons for thinking that fittingness assessment of emotions in general is different from, and can come apart from, moral assessment of emotions. Hence, if the objection to blaming people without standing is at bottom a moral objection, then accounts of moral blameworthiness in terms of whether agents would be fitting targets of moral blame can without revision capture the distinction between moral blameworthiness and standing to blame morally. On the other hand, it might be argued that morally blaming someone if you lack standing is unfitting. If this is the case, accounts of moral blameworthiness in terms of whether agents would be fitting targets of moral blame require revision. However, there is a simple revision at hand: these accounts can claim that an agent would be a fitting 'target' of moral blame when the fittingness conditions of moral blame concerning them and their φ-ing are satisfied, whether or not the blamer meets the relevant fittingness conditions. So, however the objection to blaming someone morally if you lack standing is best understood, accounts of moral blameworthiness in terms of whether agents would be fitting targets of moral blame can capture the distinction between moral blameworthiness and standing to blame morally.

2.1.3 What are Moral Excuses?

Moral excuses (and moral exemptions, which will be relevant later) can be understood in terms of their relation to the conditions of moral responsibility.[10] The conditions of moral responsibility are the conditions that must be met in order for φ-ing to be imputable to the agent in such a way that they are eligible for moral blameworthiness for φ-ing.[11] These conditions

[10] As we will see in Section 2.3, morality is not the only normative domain that admits of excuses and exemptions. Just as moral excuses and moral exemptions can be understood in terms of their relation to the conditions of moral responsibility, so too can excuses and exemptions in other normative domains be understood in terms of their relation to the conditions of responsibility in those domains. I sometimes drop the qualifier 'moral' from 'excuses' and 'exemptions' when it is obvious from the context that I mean moral excuses and moral exemptions.

[11] For an overview of contemporary debates concerning what these conditions are, see Talbert (2022). MB presupposes no particular view of what these conditions are.

are morally neutral, in the sense that an agent can meet them in performing an action whether it is morally wrong, permissible, or required. They are conditions that must be met for an agent to be *eligible* for moral blameworthiness for φ-ing. Whether an agent is *in fact* morally blameworthy for φ-ing also depends on the kind of action they perform.

An agent can fail to meet the conditions of moral responsibility in one of two ways: they may be morally exempt, or possess a moral excuse.[12] Moral exemptions show that at least a significant range of behaviour fails to be properly imputable to agents in virtue of some relatively general incapacity or impairment. Examples of moral exemptions include extreme youth, certain mental illnesses or brain disorders such as late-stage Alzheimer's disease, and somnambulism. I will sometimes use the phrase 'morally responsible agent' to refer to someone who is not exempt from moral blameworthiness. Moral excuses show that an agent fails to meet the conditions of moral responsibility with respect to some particular action or omission. Examples of moral excuses include innocent non-moral ignorance, non-negligent accidents, and duress.

With this characterisation of the conditions of moral responsibility in hand, let me make one final clarification. It might be wondered whether, in understanding moral blameworthiness in terms of whether agents would be fitting targets of moral blame, I am committed to what David Shoemaker calls a 'fitting response-dependence' account of moral blameworthiness (2017). This account is opposed to views on which the conditions of moral responsibility can be explained in terms of a principle, or set of principles, that do not essentially refer to the fittingness of moral blame. An example of such a view is that an agent meets the conditions of moral responsibility just in case the first-order desire that issues in the agent's φ-ing meshes with their second-order volitions (cf. Frankfurt 1971). According to the fitting response-

[12] For similar accounts of the distinction between moral exemptions and moral excuses, see, e.g., Strawson (1962/2003); Wallace (1994); Watson (2004: 219-259); and Sliwa (2019).

dependence account, the conditions of moral responsibility are too complex and variegated to be captured by any such principle, or set of principles. Rather, the conditions of moral responsibility can only be characterised as the conditions that must be met in order for φ-ing to be imputable to the agent in such a way that they may be a fitting target of moral blame for φ-ing (depending on the moral status of φ-ing). MB presupposes no particular view of what the conditions of moral responsibility are. It is compatible with, but does not presuppose, a fitting response-dependence account.

2.2 Some Central Platitudes Surrounding Our Deontic Moral Concepts

Platitudes, we saw in the Introduction, are statements of the judgmental and inferential dispositions that are typically had by those who possess a given concept. Let me start by explaining what notion of concept possession I have in mind, before indicating two characteristic marks of platitudes. I will then explain what it means for an analysis to 'respect' a platitude.

There is a weak sense in which someone possesses a concept, C, just in case they have propositional attitudes towards propositions that contain C as a constituent. Arguably, this is compatible with the person making various mistakes about the concept in question. To give an example of Tyler Burge's, someone might have beliefs about arthritis even though they misunderstand ARTHRITIS, not realising that arthritis is a disease only of the joints (1979). Moreover, there may be gaps in a subject's understanding of a concept, as when they do not know whether a contract must be written (Ibid.). More radically, someone might have propositional attitudes towards propositions that contain a certain concept as a constituent despite knowing almost nothing about that concept. For example, this will be the case when someone possesses the relevant concept parasitically (as an intelligent Martian might possess

a concept *via* the introduction, 'the concept expressed by the Earth term 'X''). The notion of concept possession I have in mind is a strong notion, on which possessing a concept, C, involves not only having propositional attitudes towards propositions that contain C as a constituent, but also rules out misunderstandings of concepts and gaps in understanding.[13] Platitudes, then, are statements of the judgmental and inferential dispositions that are typically had by those who possess the relevant concepts in this strong sense. From now on, when I talk about 'possessing' a concept, it is the strong sense I have in mind.

Platitudes have two characteristic marks. The first is that apparently denying a platitude associated with a concept, or making a judgment or inference that directly contradicts such a platitude, is evidence that the agent does not possess this concept (or at least is not employing it).[14] The second characteristic mark is that appearing to deny a platitude or make a directly contradictory judgment or inference tends to produce a feeling of bafflement among those who possess the relevant concept.

A couple of examples will illustrate these characteristic marks. A relatively uncontroversial example of a platitude associated with our normative concepts is that there can be no normative difference without a non-normative difference: two actions that are exactly the same in non-normative respects must be the same in normative respects (Jackson and Pettit 1995: 21). If someone apparently makes a judgment that directly contradicts this platitude – for example, by seemingly judging that action A was morally wrong and that action B was not morally wrong, while claiming that A and B were identical in non-normative respects – this counts as evidence that they do not possess the relevant concept (MORAL WRONGNESS) and will

[13] For this distinction, see, e.g., Bealer (1998: 272-275). Bealer goes on to defend a positive analysis of this strong notion of concept possession, but nothing in my argument presupposes this analysis.

[14] The qualifier 'apparently' is needed because if the seeming denial of a platitude (or the seeming making of a directly contradictory judgment or inference) rightly leads us to conclude that the agent in question does not in fact possess the relevant concept, then we cannot correctly interpret them as making judgments containing this concept.

tend to give rise to feelings of bafflement among those who possess this concept. To give a non-ethical example, if someone appears to judge that a figure is a four-sided triangle, this counts as evidence that they do not possess the concept TRIANGLE, or at least are not employing it (perhaps they thought 'triangle' meant 'square'), and those who possess this concept will tend to feel baffled.

In the Introduction, I characterised the descriptive desideratum on conceptual analyses as the condition that conceptual analyses should aim for the most part to 'respect' the platitudes surrounding the concept under analysis. Minimally, an analysis *respects* a platitude insofar as the analysis, perhaps together with further platitudes concerning its constituent parts, entails the platitude. This ensures that someone who adopted that analysis – in the sense that they possessed, and used, the concept the analysis spells out – would have the judgmental or inferential disposition stated by the entailed platitude. To give a simple example, it is a platitude surrounding AUNT that all aunts are women. The analysis, 'X is Y's aunt if and only if (Def) X is a woman sibling or sibling-in-law of one of Y's parents' respects this platitude because this analysis entails this platitude. Someone who adopted this analysis would thereby be disposed to infer from someone's being an aunt that they are a woman. An analysis might also respect a platitude by illuminating it as well as entailing it. For example, I will argue that MB respects the platitude that moral wrongs can be more or less serious. MB illuminates this platitude insofar as it explains seriousness of moral wrongdoing in terms of the clearer category of degrees of moral blameworthiness.

With this characterisation of platitudes in hand, and what it means for a conceptual analysis to respect the platitudes surrounding a concept, we can now look at some of the platitudes surrounding our deontic moral concepts. I will focus on three. As we will see, MB respects all of these platitudes.

2.2.1 Making Amends

The first platitude is a statement of an inferential disposition: *if it is morally wrong for an agent to φ, then,* ceteris paribus, *there is a reason for that agent to make amends for φ-ing without a moral excuse.* The *ceteris paribus* clause is needed for some atypical cases, for instance in which the offender is dead or, perhaps, unable to do anything that would constitute making amends (if reasons are subject to a response constraint).[15] The statement of this inferential disposition is a good candidate for a platitude surrounding MORALLY WRONG. Apparently denying the link between moral wrongness and reasons to make amends – for instance, by saying that it is rarely, or never the case that culpable moral wrongdoers need to 'put things right', 'make up for what they've done', or 'make amends' – would be evidence that the speaker does not possess MORALLY WRONG (or at least is not employing it), and would tend to produce feelings of bafflement among competent users of this concept.

MB respects this platitude by means of a bridge principle connecting the fittingness of an emotional response with reasons for acting in accordance with its motivational goal. This principle states that, if it would be fitting for an agent to feel E, then, *ceteris paribus*, that agent has a reason to fulfil E's motivational goal (cf. Skorupski 2010: 265-267). For example, if it would be fitting for an agent to fear X, then, *ceteris paribus*, that agent has a reason to become safe from X; if it would be fitting for an agent to feel grateful to Y for a good turn, then, *ceteris paribus*, that agent has a reason to thank Y and reciprocate in some way; and if it would be fitting for an agent to feel guilty about Z, then, *ceteris paribus*, that agent has a reason to make amends for Z. These reasons need not be decisive or even sufficient. For example, someone might have decisive reasons not to make amends for Z even though it would be fitting to feel

[15] See, e.g., Williams (1995a: 39).

guilty about Z, if making amends would have terrible consequences. But if guilt would genuinely be fitting, then, *ceteris paribus*, there is at least a *pro tanto* reason to make amends. Moreover, since the bridge principle is a conditional claim, it does not by itself entail, for any given emotion-type, that episodes of that type would ever be fitting; and hence does not by itself entail that there is ever a reason to fulfil the motivational goal of a given emotion-type. The *ceteris paribus* clause is needed for atypical cases. For example, perhaps if it is impossible for someone to do anything to make amends for Z, then they cannot have a reason to make amends for Z (if there is a response constraint on reasons). But it seems nonetheless that it may be fitting for them to feel guilty about Z.

It is plausible that the bridge principle states a platitude surrounding FITTING EMOTIONS. Granting that it would be fitting to feel gratitude for a good turn, for example, but denying that one has, even *ceteris paribus*, any reason to thank the relevant agent or reciprocate would be evidence that the agent does not possess FITTING GRATITUDE (or at least is not employing this concept); and so on for other fitting emotions, such as fitting fear (safety) and fitting guilt (making amends). Moreover, denial of these things would tend to produce feelings of bafflement among those who possess the relevant concepts.

Given this bridge principle, MB entails the platitude: *if it is morally wrong for an agent to φ, then,* ceteris paribus, *there is a reason for that agent to make amends for φ-ing without a moral excuse*. Moral blame, we have seen, consists in guilt, resentment, and indignation. Guilt, moreover, motivates agents to make amends. So, MB entails the platitude linking moral wrongness with reasons to make amends *via* the bridge principle.

2.2.2 More/Less Serious Moral Wrongdoing

Moral wrongs can be more or less serious. *Ceteris paribus*, it is more seriously morally wrong to torture someone than lie to them, and less seriously morally wrong to tickle them without their consent than murder them. Although there is often room for reasonable disagreement concerning which actions are more seriously morally wrong than others, the claim that there is some such distinction has the status of a platitude. Denying any distinction between more or less serious moral wrongdoing would be evidence that the speaker does not possess MORALLY WRONG, and such a denial would tend to produce feelings of bafflement among those who do possess this concept.

MB respects this platitude. Moral blameworthiness comes in degrees: it can be fitting for someone to blame someone morally more or less strongly. That is, it can be fitting for someone to feel more or less strongly indignant, resentful, or guilty about someone's φ-ing (I discuss emotional strength in relation to fittingness assessments in Chapter 5). Thus, MB entails that we can discriminate among standards on the basis of whether violation of standards without a moral excuse would be more or less morally blameworthy. Hence, MB naturally gives rise to a distinction between more or less serious moral wrongs (cf. Gibbard 1990: 45). Moreover, insofar as the notion of degrees of moral blameworthiness is intuitively clearer than that of seriousness of moral wrongdoing, MB not only entails but illuminates the distinction between more or less serious moral wrongs.[16]

[16] See Hurka (2019) for further discussion of the notion of seriousness of moral wrongdoing that links seriousness with degrees of moral blameworthiness. It might be worried that the account of seriousness of moral wrongdoing sketched above is mistaken insofar as the motives from which agents act are relevant to degrees of moral blameworthiness but not relevant to seriousness of moral wrongdoing. This is part of a more general concern about the extensional adequacy of MB, to the effect that motives are differentially relevant to moral wrongness and moral blameworthiness. I address this extensional challenge in Chapter 4.

2.2.3 Moral Responsibility and Moral Agency

The next platitude I will discuss concerns the link between moral responsibility and what I will call 'moral agency'. A moral agent is an agent whose actions – and, perhaps, beliefs, emotions, intentions, and desires – can aptly be assessed as morally wrong, permissible, or required. (I will sometimes call a moral agent an 'agent who is subject to moral requirements'). It seems plausible that moral responsibility places some kind of constraint on moral agency. For example, Michael Smith writes:

> Roughly speaking, those normative claims that entail the possibility of holding some agent responsible are deontic… …Thus, for example, when people do something that they shouldn't do, it follows more or less immediately that they are candidates for being held responsible. There may be exemptions or excuses, so we may not hold them responsible in fact, but they are at least candidates. (2005: 10)

Smith speaks here of deontic claims in general, but our focus is more specifically on deontic moral claims: applied to this specific case, the relevant claim is that moral responsibility places some kind of constraint on being an apt target of deontic moral assessment. While this idea is intuitively compelling, Smith's particular formulation of it requires some adjustments. Smith claims that the link concerns whether agents are 'candidates' for being held (morally) responsible, and he connects this notion of candidacy with (moral) excuses and (moral) exemptions. Someone is a 'candidate' for being held morally responsible, Smith seems to suggest, just in case, unless they have a moral excuse or exemption, it would be fitting or appropriate to hold them morally responsible. But while it is plausible that moral excuses do not show that the relevant agents were not subject to the relevant moral requirements, this claim is much less plausible when applied to moral exemptions.

Although some have taken excuses to undermine moral wrongness as well as moral blameworthiness, it seems more faithful to how these concepts are ordinarily understood to claim that excuses show that an agent is not (fully) morally blameworthy for φ-ing while leaving the moral wrongness of φ-ing intact (Sliwa 2019: 42-43). For example, suppose that Adam accidentally and non-negligently breaks his friend's precious vase. Provided that Adam is not exempt from moral blameworthiness, it seems plausible that his action is morally wrong. After all, the fact that Adam's action was accidental does not provide a justification for it. So, defenders of the view that moral wrongness is analysable in terms of moral blameworthiness had better incorporate moral excuses into their analysis.[17]

Moral exemptions are a different matter. Suppose that Tom, a two-year-old, hits his mother in anger after she takes away his toy. Tom is exempt from moral blameworthiness. He lacks moral concepts, and many of his general agential capacities, such as capacities for deliberation and self-control, are underdeveloped. It also seems intuitive that he does not act morally wrongly in hitting his mother. Generally, he is not subject to moral requirements. This is not to deny that it can make sense to blame young children and impute wrongdoing to them proleptically, as a way of helping them to become morally responsible agents and moral agents. But these activities tend to be compresent, it seems, with the thought that young children are not *really* morally blameworthy and are not *really* subject to moral requirements. To take a different example, suppose that while Jenny is sleepwalking she strikes her partner. Jenny is exempt from moral blameworthiness for what she does while somnambulistic. Moreover, intuitively she does not act morally wrongly in striking her partner. Generally, Jenny is not subject to moral requirements while sleepwalking.

[17] Given that moral excuses have been defined solely in terms of their relation to the conditions of moral responsibility, incorporating moral excuses does not make MB circular.

The claim that agents who are exempt from moral blameworthiness are not moral agents is supported not only by intuitions about particular cases but also more generally by links between moral blameworthiness, moral requirements, and legitimate demands. A demand is legitimate when the demander has the authority to make it of the demandee. We can challenge the legitimacy of a demand by saying things like, 'That's not something you can demand/expect of me'. A demand can be legitimate without being all-things-considered morally or practically justified. For example, it is often unjustified to demand that someone φ if there is no indication that they will not φ, but this need not undermine the legitimacy of the demand. Imagine that Sophie promises Chris to help him move house, and Chris demands that Sophie keep her promise even though all signs point to her already having a firm intention to do so. Sophie might well reply, 'I was going to keep my promise anyway!'. This challenges the justification for Chris's demand, but not its legitimacy.

We can appeal to links between moral blameworthiness, moral requirements, and legitimate demands to argue on the following lines for the claim that agents who are exempt from moral blameworthiness are not subject to moral requirements.[18] First,

> (P1) If an agent is exempt from moral blameworthiness, it would not be legitimate for anyone to demand that they comply with any moral requirements.

Support for this claim comes from reflection on what distinguishes demands from other speech acts, such as requests. Plausibly, demanding that someone performs an action, unlike requesting that they perform it, involves conveying that if they fail to perform it they will be blameworthy if they lack a good excuse. This would seem illegitimate if the agent is exempt from the relevant kind of blameworthiness. Second,

[18] Strawson (1962/2003: 86) hints at an argument along these lines. The claim that there are tight (conceptual) links between moral blameworthiness, moral requirements, and legitimate demands is an important theme of Darwall (2006a).

> (P2) If it would not be legitimate for anyone to demand that an agent complies with any moral requirements, they are not subject to any moral requirements.

This claim is *prima facie* plausible. On the face of it, moral requirements have a tight connection with legitimate interpersonal demands, such that it is intuitively plausible that an agent who is not a candidate for legitimate interpersonal demands for compliance with moral requirements is not subject to any moral requirements – they are not a moral agent.

Against P2, it might be argued that not all moral requirements are such that it would be legitimate for at least some agents to demand compliance with them. For example, John Skorupski writes:

> There is, importantly, a *difference* between what we have a moral obligation to do and what anyone, including the 'community', can permissibly demand of us. (Consider obligations of gratitude, of loyalty in friendship, of tact; moral obligations to oneself, if there are such.)[19] (2010: 375)

(Skorupski talks in this passage about *permissible* demands, whereas our focus is on *legitimate* demands. Still, the same point might be thought to apply in both cases). Now, we might doubt that there really are moral obligations of these kinds.[20] But even if there are, they do not undermine P2. P2 claims that if it would not be legitimate for anyone to demand that an agent complies with *any* moral requirements, they are not subject to any moral requirements. The most central and important moral requirements – those which forbid serious harm against others, for instance – are such that it would be legitimate for others (perhaps, anyone) to demand compliance. It is not clear that we can make good sense of the possibility of someone not being subject to these central moral requirements while still being subject to the more

[19] Cf. Manela 2015 on gratitude.
[20] See Wellman (1999) for critique of the claim that there are moral obligations of gratitude.

marginal examples of moral requirements that Skorupski mentions. (How, for example, could someone be morally required to be tactful to someone without being morally required not to murder them, given the relative importance of the interests at stake?). So, P2 is still very plausible even if there are some marginal examples of moral requirements with which no-one can legitimately demand compliance.

From P1 and P2, it follows that if an agent is exempt from moral blameworthiness, they are not subject to any moral requirements. Let me label this conditional claim as follows:

> *The Moral Responsibility-Moral Agency Link (RAL)*: If an agent is exempt from moral blameworthiness, they are not a moral agent.[21]

RAL, like Smith's remarks quoted above, aims to capture the idea that moral responsibility places some kind of constraint on being subject to deontic moral assessment, but it does so in a more plausible way. Moral excuses do not show that the relevant agents were not subject to the relevant moral requirements, but moral exemptions do show this.

RAL is another platitude surrounding our deontic moral concepts that MB is well-placed to respect. The status of RAL as a platitude is borne out both by our reactions to particular cases (such as the examples of Tom and Jenny given above) and the more general argument linking moral blameworthiness, moral requirements, and legitimate demands. Failing to appreciate that these notions are connected by the inferential patterns described in P1 and P2 would be evidence of not possessing the relevant concepts. If someone took himself to be

[21] The claim that agents who are exempt from moral blameworthiness are not subject to moral requirements is accepted, sometimes implicitly, by Strawson (1962/2003: 86); Watson (2004: 224-225); Darwall (2006a: 14); and Sliwa (2019: 70-71). This claim needs to be relativized to particular time-ranges. Suppose that an agent, A, suffers from a severe cognitive impairment from $t_0 - t_1$ but not from $t_1 - t_2$. Then it may be that A is exempt from moral blameworthiness from $t_0 - t_1$ but subject to moral requirements from $t_1 - t_2$. So the claim that agents who are exempt from moral blameworthiness are not subject to moral requirements needs to be understood as asserting that if agents are exempt from moral blameworthiness *during a given time-range*, they are exempt from moral requirements *during that time-range*. For readability, I omit this qualification in what follows. The conditionals below connecting moral blameworthiness to legitimate demands, and legitimate demands to moral requirements, also need to be understood as relativized to particular time-ranges.

in a position legitimately to demand that agents who are exempt from moral blameworthiness – Alzheimer's patients, somnambulistic agents, very young children, and the like – comply with moral requirements, that would be evidence that he does not possess these concepts (LEGITIMATE DEMAND, MORAL REQUIREMENT, MORAL BLAMEWORTHINESS). Moreover, those who possess these concepts would tend to feel baffled by such an agent. And the same would also hold for someone who accepted that some people are subject to moral requirements but denied that it would ever be legitimate for anyone to demand that moral agents comply with their moral requirements.

MB entails RAL. This is because MB does not admit exemptions into its analysis of moral wrongness. Agents who are exempt from moral blameworthiness never violate standards such that, if they violated those standards without a moral excuse, they would be morally blameworthy for violating them. (Since the closest possible worlds in which they violate these standards are ones in which they are exempt from moral blameworthiness). Given this, MB entails RAL.

2.3 MB as a Metaphysical Analysis

Moral requirements are not the only kinds of requirements. For instance, there are prudential requirements, legal requirements, epistemic requirements, and requirements of etiquette. My argument for MB as a metaphysical analysis is that it fits into an attractive general pattern of analysis of requirements across different normative domains, in which different kinds of requirements are analysed in terms of their connections with different kinds of negative reactions. This lends support to MB as a metaphysical analysis in two ways. First, we can appeal to this general pattern of analysis to explain what unifies different kinds of requirements.

Second, insofar as analogous views to MB hold in other normative domains, we should expect MB to hold in the moral domain as well.

Generalising from MB suggests the following pattern of analysis for different kinds of requirements:

> *Requirement*: For an agent to be D-required to φ is for not φ-ing to violate standards such that, if the agent violated those standards without a D-excuse, they would be worthy of D-negative reactions for violating those standards.

'D-required' means required by the lights of some normative domain, such as morality, prudence, or the law; 'D-negative reactions' means the negative reactions distinctive of some normative domain; and 'D-excuse' means excused by the lights of some normative domain. (As we will see shortly, this pattern of analysis may need to be modified slightly for some kinds of requirements, such as legal obligations). I will now explain how various kinds of requirements might be analysed in terms of this pattern. The discussions will be much briefer than my discussion of MB, and I do not take myself to establish any of these analyses of non-moral requirements conclusively. But the discussions should suffice to show that the pattern of analysis provided by Requirement is very promising.

I will start by discussing prudential requirements. Applying the pattern of analysis given in Requirement to prudential requirements yields:

> *Prudential Requirement*: For an agent to be prudentially required to φ is for not φ-ing to violate standards such that, if the agent violated those standards without a prudential excuse, they would be prudentially blameworthy for violating them, where 'φ' stands for an object of prudential assessment.

As in the case of MB, Prudential Requirement is not in competition with accounts of which actions are prudentially required and why. Quite what counts as excusing someone from

prudential blame is an issue that I will leave open, but many conditions that provide moral excuses seem also to provide prudential excuses. For instance, someone might not be prudentially blameworthy because of non-culpable non-evaluative ignorance, or because their imprudent action was accidental.

To fill out Prudential Requirement, we need an account of prudential blame. This account, moreover, must spell out prudential blame independently of prudential requirements, on pain of rendering Prudential Requirement circular. In principle, any account of prudential blame that meets this independence constraint might be slotted into Prudential Requirement, but I will develop a specific proposal that focusses on feeling frustrated with someone – either yourself or someone else. This is plausibly a blaming reaction, unlike, say, feeling sad or worried. Feeling frustrated with yourself certainly seems to be a common reaction in the wake of violating prudential requirements. Moreover, it is common to feel frustrated with people you care for when they act in imprudent ways. For example, you might feel frustrated with a friend who gets drunk the night before an important interview and blows their chances of getting the job. Of course, we can also feel frustrated with people who have not behaved imprudently – for instance, because they have been inconsiderate. Furthermore, we can feel frustrated with ourselves even when we have not violated a prudential requirement. For example, you might feel frustrated with yourself for missing an easy shot in a tennis game. But we can home in on a distinctly prudential kind of blame by focussing on being frustrated with someone *as* someone who cares for that person for their own sake.[22] (Since you can care for yourself, you can be frustrated with yourself *as* someone who cares for yourself for your own sake). Prudentially blaming someone for φ-ing, then, consists in feeling frustrated with them for φ-ing *as* someone

[22] It might be worried that appealing to caring in an account of prudential blame will make it viciously circular to analyse prudential requirements in terms of prudential blameworthiness. Part of caring about someone for their own sake, we might think, is wanting them to fulfil their prudential requirements. I cannot discuss this objection in detail here, but there are plausible accounts of caring for someone for their own sake that make no essential appeal to prudential concepts or judgments. See, e.g., Rowland (2019: 82-83).

who cares for them for their own sake.[23] Prudential blame, so understood, is distinct from moral blame directed at morally culpable prudential failures. We can appeal to this account of prudential blame to fill out Prudential Requirement.

Next, let me consider epistemic requirements. Applying the pattern of analysis given in Requirement to epistemic requirements yields:

> *Epistemic Requirement*: For an agent to be epistemically required to φ is for not φ-ing to violate standards such that, if the agent violated those standards without an epistemic excuse, they would be epistemically blameworthy for violating them, where 'φ' stands for an object of epistemic assessment.

As with MB and Prudential Requirement, Epistemic Requirement is not in competition with first-order theories of which φ-ings are epistemically required and why. Moreover, I will leave open which conditions excuse agents from epistemic blame.

To fill out Epistemic Requirement, we need an account of epistemic blame. The account, moreover, needs to explain epistemic blame independently of epistemic requirement, or else Epistemic Requirement will be circular. I will draw on Brian McElwee's (2017: 511-515) account of epistemic blame, but in principle any account of epistemic blame that meets this independence constraint could be plugged in. McElwee explains epistemic blame in terms of excluding people from the epistemic community, where this involves such things as not relying on their testimony and excluding them from communal inquiry. For example, forming beliefs too easily – believing that *p* when the evidence for *p* is weak, say, or when there is strong countervailing evidence – merits exclusion of this kind (2017: 513). As an illustration, if Joe

[23] McElwee (2017: 509-510) proposes a different understanding of the negative reactions distinctive of the prudential domain. He focusses on the charge of foolishness and the attitudes that tend to stand behind this charge, which he takes to include loss of esteem, contempt, and distaste. However, while we can fittingly respond to imprudence in some of these ways, these negative reactions seem primarily to reflect assessments of prudential *vice* rather than *impermission*.

believes that some water is drinkable on the basis of weak evidence – for instance, that it is reasonably clear – then not relying on his testimony (at least on water-related matters) is merited, at least until he has shown himself to be epistemically trustworthy. Epistemic blame, understood in this way, is distinct from moral blame directed at morally culpable epistemic failures (2017: 512-513). We can appeal to this account of epistemic blame to flesh out Epistemic Requirement.

Let me turn finally to legal and other institutional obligations. Analysing such obligations in terms of sanctions seems promising. Sanctions figure prominently in H. L. A. Hart's classic analysis of legal obligation (1961/2012). Hart aims to explain legal obligations as a subset of social rules. In Hart's view, social rules are distinguished from mere habits by three features. First, members of a social group that accepts a rule use it to guide their conduct. Second, they take criticism of those who deviate from it to be merited. Third and finally, they take it to be legitimate to demand that others conform their conduct to it (55-60). Very roughly, Hart holds that rules impose legal obligations when deviation is taken to merit (typically, official) sanctions (86-87, 97-98). Hart's analysis of legal obligation is controversial, and I do not mean to commit myself to all its details. But insofar as there is plausibility in an approach to understanding legal obligation that is broadly Hartian in its focus on sanctions that are taken to be merited or fitting, then it seems that an analogous view to MB holds in the legal domain. Admittedly, the Hartian analysis of legal obligation departs from the pattern described above. This is because it analyses legal obligations in terms of negative reactions that are *taken to be* merited, rather than negative reactions that are *actually* merited.[24] But the common focus on negative reactions secures a close analogy between these analyses.

[24] Is this pattern of analysis plausible for other kinds of requirements besides legal obligations, and if so when? Other kinds of institutional obligations, such as the obligations one might have as the member of a club, are promising candidates. Perhaps requirements of etiquette are also best analysed in terms of negative reactions that are taken to be fitting, as opposed to negative reactions that are actually fitting. What all of these requirements

By conforming to an attractive general pattern of analysis of different kinds of requirements, permissibility, and wrongness across different normative domains, MB forms part of an attractive explanation of what unifies different kinds of requirements. Moreover, insofar as analogous views to MB hold in other normative domains, this should bolster our confidence that MB holds in the moral domain as well. To show why these are strengths of MB, it will help to contrast MB with some rival views of the structure of the moral domain. We saw in the Introduction that, provided MB is extensionally adequate, there is one main rival family of views to MB. On these views, the moral wrongness relation is distinct from the moral blameworthiness relation, and facts about moral wrongness explain facts about whether φ-ing violates standards such that, if an agent violated those standards without a moral excuse, they would be morally blameworthy for violating them.

There are different ways of developing these views. One way is to analyse the moral wrongness relation by drawing on what are usually interpreted as first-order moral theories. Thus, we might claim that for it to be morally wrong for an agent to φ is for φ-ing to fail to have the best consequences, or for φ-ing to be disallowed by principles that no-one could reasonably reject, and so on. The trouble with this approach is that it leaves it unclear what unifies different kinds of requirements. While it seems clear enough that contractualists can explain how there are sometimes *moral* requirements to act prudently and cultivate certain epistemic practices, it is obscure how we could give contractualist accounts of distinctly prudential and epistemic requirements. Perhaps views analogous to moral consequentialism might be defended in the prudential and epistemic domains, by identifying distinctive prudential and epistemic goods. But what about legal and other institutional obligations? These too are kinds of requirements, but consequentialist accounts of them look like non-starters.

---

seem to have in common is that for them to exist there must be a social practice of acceptance or recognition of them.

Insofar as MB fits into a general pattern of analysis that promises to explain what all of these kinds of requirements have in common, it is more attractive to defend MB than to adopt what are usually interpreted as first-order moral theories as metaphysical analyses. Moreover, views such as consequentialism and contractualism are harder to defend as metaphysical analyses than MB because they have many more competitors. (As we will see in Chapter 4, MB is largely neutral with respect to contentious first-order debates).

Instead of analysing the moral wrongness relation by drawing on what are usually seen as first-order moral theories, someone might instead defend a rival analysis that operates at a similar level of abstraction to MB. This is the approach taken by Victor Tadros (2016: 27-46), Richard Rowland (2019: 171-192), and Kieran Setiya (2022).

According to Tadros, for it to be morally wrong for an agent to φ is for φ-ing to be disfavoured by sufficiently significant reasons that φ-ing is not valuable as a free expression of agency. He summarises this account, which he calls the 'Respecting Value' account of moral wrongness, as follows:

> Certain actions are not valuable as free expressions of agency because of the significance of the reasons against acting in that way. Respect for these values makes valuing one's free performance of such actions inappropriate. This explains the idea that these values constrain. And this explains wrongness. (2016: 41)

To illustrate, consider the following pair of examples. Suppose that a writer spoils a novel that could have been great by giving it a bad ending (41-42). Although the novel is not as good a novel as it could have been, in Tadros's view this does not undermine its value as a free expression of agency, and therefore giving it a bad ending is not morally wrong. Contrast this with a case of morally wrongful assault. Here, the values that the agent disrespects in assaulting

someone undermine the value of their action as a free expression of agency, and therefore assaulting them is morally wrong (Ibid.).

One worry about Tadros's account is that the evaluative category he centrally appeals to – value *as* a free expression of agency – seems obscure. This is a kind of attributive goodness: goodness as or *qua* a certain kind of thing. Not every kind of thing is such that it is possible to be good as a member of that kind. For instance, there is no such property as being a good pebble or smudge or cloud (Thomson 2007: 22). While value as a free expression of agency is perhaps more acceptable than these examples, it still seems obscure what could make something valuable in this way. (Contrast this putative form of attributive goodness with being a good bicycle or a good knife, which are much clearer categories).

Moreover, it is difficult to see how Respecting Value could fit into an attractive general pattern of analysis that explains what different kinds of impermission have in common. In the case of epistemic impermission, for instance, it would seem that the analogous evaluative category (as applied to beliefs) would be something like, 'valuable as a free expression of belief-forming capacities'. The idea would then be that there are certain reasons against beliefs that are sufficiently significant that believing against these reasons is not valuable as a free expression of belief-forming capacities. But this evaluative category also seems obscure. More significantly, the resulting account of epistemic impermission is extensionally inadequate. Believing that *p* can be epistemically impermissible even if there are no reasons against believing that *p*, if the reasons for believing that *p* are very weak. Finally, it is difficult to see how to give an analysis of prudential impermission that is analogous to the *Respecting Value* account of moral wrongness.

According to Rowland, for an agent to be morally required to ϕ is 'for [that agent] to have sufficient reason to ϕ and for there to be sufficient non-role-dependent reason for [that

agent] to have (non-instrumental) pro-attitudes towards her making amends for not ϕ-ing if she does not ϕ' (2019: 176). Rowland calls this the 'Amends-Based Buck-Passing' account. The restriction to *non-role-dependent reasons* for making amends is significant, because we can have reasons for making amends for actions in virtue of occupying certain roles. For example, perhaps Tim might have reasons for making amends with Susan for not making an effort to keep in touch with her in virtue of his being friends with her. Such a failing would seem not to be morally wrong. In the case of morally wrong actions, however, our reasons for making amends seem not to be role-dependent (Rowland 2019: 176). For example, someone who seriously harms another when there is no good reason for doing so has reason to make amends regardless of what role(s) they occupy (Ibid.).

One difficulty with Rowland's account concerns making amends. As we will see in Chapter 5, it is crucially important to making amends that, in making amends for φ-ing by apologising, offering compensation, etc., we express our guilt-feelings, along with our conviction that resentment and indignation on the part of others would be fitting. Without these expressive qualities, making amends could not have many of its reparative effects. This strongly suggests that it is a mistake to analyse moral requirements in terms of norms on making amends (or for having pro-attitudes towards making amends) while leaving out of the analysis the emotions of which making amends is an expression.

Moreover, Amends-Based Buck-Passing fares less well than MB with respect to explaining what unifies different kinds of requirements. Rowland argues that a defender of Amends-Based Buck-Passing can hold that non-moral requirements are analysable in terms of *role-dependent* reasons for action and *role-dependent* reasons for making amends (2019: 180). For instance, on this approach, for Katherine to be under a requirement of friendship to water Paulina's plants while she is away is for Katherine to (A) have sufficient reasons to water Paulina's plants partly in virtue of being friends with Paulina, and (B) have sufficient reasons

to have (non-instrumental) pro-attitudes towards her making amends for not ϕ-ing if she does not ϕ partly in virtue of being friends with Paulina. This is a plausible result: practices of making amends seem just as home in the context of friendship as in the context of morality. Moreover, it is also plausible that it is partly in virtue of being friends with someone that we have reasons to make amends with them when we violate requirements of friendship to them (and reasons not to violate such requirements in the first place).

However, this strategy of explaining what unifies different kinds of requirements is less successful with respect to non-moral requirements that are not role-dependent and that are not closely tied with making amends. Moreover, MB is able to offer structurally similar accounts of role-dependent obligations like requirements of friendship, so Amends-Based Buck-Passing does not have an advantage over MB in this respect.

Prudential requirements and epistemic requirements are not, or at least need not be, role-dependent. Someone may be under a prudential requirement to φ, or an epistemic requirement to believe *p*, where this has nothing to do with any of the roles they occupy. So, appealing to role-dependent reasons to analyse these kinds of requirements seems unpromising. Hence, it seems that a defender of Amends-Based Buck-Passing is not well-positioned to explain what all of these different kinds of requirements have in common. In response, it might be suggested that a defender of this view might try to draw a (non-circular) distinction between moral amends and other kinds of amends, such as prudential and epistemic amends, and then appeal to these different kinds of amends to explain the similarities and differences between moral, prudential, and epistemic requirements. However, it is not clear that good sense can be made of these putative other kinds of amends, especially in the prudential case. Someone who violates a prudential requirement may have reasons to change their prudential policies, but this does not seem to be well-understood as a kind of amends, given that they have reasons to

change these policies *anyway*, irrespective of them leading him or her to violate this particular prudential requirement.

Moreover, a defender of MB might analyse role-dependent obligations like requirements of friendship in terms of whether guilt and resentment would be fitting partly in virtue of the relevant parties occupying the relevant roles. In the case of requirements of friendship, this approach yields:

> *Requirements of Friendship*: For an agent, A, to be required to φ as a friend of B's is for not φ-ing to violate standards such that, if A violated those standards without an adequate excuse, it would be fitting for A to feel guilty for violating them partly in virtue of A being B's friend, and for B to resent A for violating them partly in virtue of B being A's friend.[25]

Requirements of Friendship is just as attractive an analysis of requirements of friendship as Rowland's analysis in terms of role-dependent reasons for action and pro-attitudes towards making amends. Guilt and resentment seem just as home in the case of unexcused violations of requirements of friendship as moral requirements. Moreover, insofar as guilt and resentment are closely linked with making amends, Requirements of Friendship accounts for the close tie between requirements of friendship and making amends.

According to Setiya, moral wrongness is analysable partly in terms of moral rights. According to:

[25] A difference between this analysis and MB, Prudential Requirement, and Epistemic Requirement is that, whereas these latter analyses are formulated specifically in terms of moral excuses, prudential excuses, and epistemic excuses, Requirements of Friendship is formulated in terms of 'adequate excuses'. Presumably, given that the blaming reactions at issue overlap significantly with moral blaming reactions (the difference being, in the case of moral blame, indignation), the same kinds of conditions that would excuse agents from moral blame when they have violated a moral requirement would also excuse them from guilt and resentment when they have violated a requirement of friendship.

> *Moral Wrongness as the Absence of a Right*: For an action to be morally wrong is for it to be something one should not do that one has no right to do – if others could simply prevent one from doing it, that would not infringe one's rights. (1124)

The rights in question are claim-rights, rather than liberty-rights (1126). The background assumption is that agents have a general claim-right against others not to interfere directly with their agency, but this right is not unlimited (1124). When agents have no such right, and when they should not perform the action in question, then it is morally wrong.

Moral Wrongness as the Absence of a Right, like any analysis of moral wrongness, faces its share of challenges. In particular, it is reliant on the availability of an account of moral claim-rights that is suitably independent of moral wrongness (Setiya 2022: 1125). The difficulty I will focus on is that it is hard to see how to fit Moral Wrongness as the Absence of a Right into an attractive general pattern of analysis that unifies different kinds of requirements. Many requirements, such as epistemic requirements and prudential requirements, have little or nothing to do with rights. So, a cost of accepting Moral Wrongness as the Absence of a Right is that it leaves us unable to explain what all of these different kinds of requirements have in common.

Comparison of MB with Respecting Value, Amends-Based Buck-Passing, and Moral Wrongness as the Absence of a Right shows that it is a significant virtue of MB as a metaphysical analysis that it puts us in a good position to give an attractive general account of what unifies different kinds of requirements.

2.4 Conclusion

This was the first of two chapters giving my main positive arguments for MB. I explained the various elements of MB, argued that MB satisfies the descriptive desideratum on conceptual

analyses explained and defended in the Introduction, and argued that it provides a plausible analysis of deontic moral relations as well as concepts. In the next chapter, I argue that MB meets the prescriptive desideratum on conceptual analyses.

Chapter 3

Arguments for Moral Wrongness as Moral Blameworthiness: Part 2

In Chapter 2, I explained how MB respects, and provides plausible interpretations of, central aspects of how we ordinarily use and understand deontic moral concepts. This chapter argues that the account of deontic moral concepts provided by MB explains how deontic moral concepts serve two valuable social functions: what I call 'the Deliberative Reliability Function' and 'the Respect Function'. Together, the arguments of Chapter 2 and this chapter show that MB meets both the descriptive and prescriptive desiderata on successful conceptual analyses that were explained and defended in the Introduction. MB provides analyses of deontic moral concepts that are broadly in line with the deontic moral concepts we in fact use. At the same time, MB articulates deontic moral concepts that are worth keeping.

Section 1 argues that the account of deontic moral concepts provided by MB explains how deontic moral concepts serve the Deliberative Reliability Function. This is the function of protecting important interests through encouraging reliable patterns of deliberation and motivation. Examples of these interests include the interests we have in not being killed, hurt, or lied to (and the further interests we have in being able to rely on not being treated in these ways). The argument of this section develops the claim that other-directed moral blaming emotions, together with their expression in overt moral blame, can act as 'proleptic mechanisms': we treat people as though they have strong normative reasons not to act as we blame them for acting, and as a result of this they come to be appropriately motivated on relevantly similar future occasions (cf. Williams 1995a: 40-44; Fricker 2016, 2018; Bagley 2017; Tsai 2017; McGeer 2019). Section 1 also explains some of the background assumptions

my argument makes about emotions' responsiveness to fittingness judgments, and what I mean by 'function'.

Section 2 argues that the account of deontic moral concepts provided by MB explains how deontic moral concepts serve the Respect Function. This is the function of providing social recognition of the dignity of persons and their worthiness of respect. The argument of this section develops the claim that making amends expresses respect for persons. Respect for persons is conceptually connected with the dignity of persons: if someone is worthy of respect as a person, they have dignity as a person, and *vice versa*. Insofar as a practice of using deontic moral concepts tends fairly reliably to produce the effect that wrongdoers make amends for their misdeeds, this practice tends fairly reliably to provide social recognition of the dignity of persons and their worthiness of respect.

3.1 The Deliberative Reliability Function

I will start by explaining some of the background claims about emotions' responsiveness to fittingness judgments on which my argument will rely (these claims were laid out more fully in 1.2.2 of the Introduction), and what I mean by 'function'.

Our emotions are significantly, though imperfectly, responsive to our fittingness judgments about them – much more so than they are responsive to other kinds of normative judgments we make about them, such as judgments about their prudential value (D'Arms 2005).[1] According to MB, deontic moral judgments are in large part judgments about the

[1] Moreover, our emotions are to some extent directly rationally sensitive to fit-making facts. For example, fear at an obvious and immediate threat will typically not be mediated by a judgment as to whether fear would be fitting. Rather, fear in such cases is more likely to be a direct response to the facts in virtue of which it is fitting. My reason for focussing on the sensitivity of fittingness judgments in particular is to explore one central way in which deontic moral concepts are valuable. Deontic moral concepts feature in deontic moral judgments, and deontic moral judgments serve to regulate our dispositions towards moral blame. They regulate our dispositions towards moral blame because, according to MB, deontic moral judgments just are (in large part) judgments about the

fittingness of moral blaming emotions. Given the responsiveness of our emotions (including our moral blaming emotions) to our fittingness judgments, this means that deploying deontic moral concepts will, over time, tend to bring about a loose harmony between our deontic moral judgments and our dispositions towards moral blame, such that we are, e.g., disposed to feel indignation towards, and only towards (more or less), what we judge to be unexcused moral wrongs. I will argue that deontic moral concepts are valuable insofar as we can in this way regulate our dispositions towards moral blame by deploying them. The value of such regulation is explained, in turn, by the value of moral blaming emotions.

The fact that emotions are only *significantly* responsive to our fittingness judgments, rather than *perfectly* responsive, might at first seem a limitation of the value of deontic moral concepts as MB understands them. But this feature of emotions will in fact be very important to my defence of the value of these concepts. As has often been noted, the fact that emotions can arise and persist in the face of contrary judgments concerning their fittingness helps to explain one way in which they can be epistemically valuable (e.g., D'Arms 2005: 9-10; Tappolet 2016: 167-168). Emotions can be epistemically valuable in prompting us to reflect critically on our normative judgments when they conflict with them. To return to an example given in the Introduction, someone who has been brought up in a racist society may be led to question their racist views by persistent, recalcitrant guilt-feelings concerning their treatment of racial minorities. As we will see shortly, it is an important fact about guilt-feelings that people often experience them in response to others' moral blame even when they initially judge themselves not to have acted culpably morally wrongly. Recalcitrant guilt-feelings caused by moral blame can have significant epistemic value in leading blamees to reflect critically on the moral judgments they accept.

---

fittingness of moral blame, and our dispositions towards moral blame are significantly responsive to our fittingness judgments about moral blame.

I will argue in this section and the next that deontic moral concepts serve two valuable social *functions*. The notion of function I have in mind is the following. For a concept to have the function of producing some effect is for use of it reliably (even if not without exception) to produce that effect under propitious circumstances; and, moreover, for the effect to be in some important way useful or valuable.[2] There is therefore no implication that we came to possess the relevant concepts *because of* their ability to serve these functions, whether by evolutionary or cultural pressures, or by conscious human design.[3] In arguing that deontic moral concepts serve valuable social functions, then, I have two tasks. The first is to show that using deontic moral concepts does in fact reliably produce the relevant effects. The second is to show that the relevant effects are valuable.

The first valuable social function served by deontic moral concepts is the Deliberative Reliability Function:

> *The Deliberative Reliability Function*: Use of deontic moral concepts reliably produces the effect of protecting important interests through encouraging reliable patterns of deliberation and motivation, such as the interests we have in not being killed, hurt, or lied to (and the further interests we have in being able to rely on not being treated in these ways).[4]

---

[2] See Queloz (2022: 1261) for a similar notion of conceptual function. A difference is that Queloz's notion focusses on whether concepts tend to contribute to the satisfaction of what he calls concept users' 'concerns', which comprises their 'needs, interests, desires, projects, aims and aspirations'. In contrast, my notion of conceptual function leaves open the possibility that a concept may serve a function because of its valuable effects, even if that value is not reflected in the present concerns of concept users.

[3] It is worth emphasising that my use of 'function' is stipulative. There are many uses of 'function' on which to say that something has the function of producing an effect does have these kinds of implications. (For example, when it is said that the function of hearts is to pump blood).

[4] See Williams (1985: 182-187; 1995c: 205) for similar claims. Williams himself traces the core idea back to Hume (1995b: 205). It may seem odd to draw on Williams in defending an account of the value of deontic moral concepts, given his opposition to what he calls 'the morality system' (1985: Chapter 10). However, although Williams claims that we would be better off without the notion of 'moral obligation', he uses this as a term of art for a particularly demanding conception of obligation, and argues that the concept of obligation, when freed from this demanding conception of it, is a valuable one. When he explains what is distinctive about the concept of obligation, he often appeals to its connection with blame (1985: 186-188). It is, I think, accurate to claim that

Central to my argument that deontic moral concepts serve this function will be the idea that other-directed moral blaming emotions, together with their expression in overt moral blame, can act as 'proleptic mechanisms': we treat people as though they have strong normative reasons not to act as we blame them for acting, and as a result of this they come to be appropriately motivated on relevantly similar future occasions (cf. Williams 1995a: 40-44; Fricker 2016, 2018; Bagley 2017; Tsai 2017; McGeer 2019). (I will explain shortly the sense in which morally blaming someone for φ-ing treats them as if they have significant normative reasons not to φ).

Bernard Williams called this a 'proleptic mechanism' because of his views in the theory of normative reasons: in his view, very roughly, an agent has a normative reason to φ if and only if they could be motivated to φ as a result of fully informed, procedurally rational deliberation from their present set of motives.[5] (Views of this kind, which posit a significant link between an agent's reasons and their motives, are sometimes called 'internalist').[6] For Williams, then, blaming people can bring it about *that they have* certain reasons as a result of treating them *as if* they have those reasons; hence the label 'proleptic mechanism'. I will remain neutral as to whether some form of internalism in the theory of normative reasons is correct. However, I will stick with the label 'proleptic mechanism', given its familiarity in the literature, even though it is strictly speaking inapt if internalism is false.

In some cases, we morally blame agents who already care significantly about the relevant moral considerations, but which caring was, for whatever reason, not effective in getting them to avoid the relevant action on this occasion. In such cases, moral blame serves as a forceful reminder of the importance of certain considerations that the blamed agent was

Williams does not think we would be better off without deontic moral concepts (as understood by MB) entirely, but only that we would be better off without a particularly demanding conception of them.

[5] See Williams (1981: 101-113, 1995a: 35-45, and 2001). Williams makes clear that he takes this to be a necessary *and* sufficient condition for having a reason at (1995a: 35-36).

[6] See Finlay and Schroeder (2017) for an overview.

antecedently disposed to see as important. However, in other cases, moral blamees do not already care significantly about the relevant moral considerations. It is in these cases that moral blame can act as a proleptic mechanism: by morally blaming someone for φ-ing, and hence treating them as if they have strong normative reasons not to φ, they can come to be strongly motivated to avoid that kind of action on relevantly similar future occasions.

There are different ways in which moral blame can act as a proleptic mechanism. Sometimes, a practice of morally blaming agents for φ-ing generates desires not to φ that are not closely connected to the (presumed) moral wrongness of φ-ing. For example, suppose that Tony could either keep his promise to help Fiona or stay home and watch his favourite T.V. show. Now suppose that, if he stays home, Fiona and other agents in his social world will morally blame Tony. This changes the options that are open to him. His choice is not simply between helping Fiona and staying home, but rather between, on the one hand, helping Fiona, and, on the other hand, staying home *and getting blamed for this*. Insofar as Tony does not care directly about keeping his promise to Fiona, but is motivated to avoid blame (for instance, because he finds it unpleasant), then blaming him can, in principle, produce an instrumental desire to keep his promise, because it brings it about that helping Fiona is instrumental to avoiding blame. But this desire is not closely connected to the moral wrongness of breaking his promise to help Fiona – in particular, it has little or nothing to do with the considerations in virtue of which this is morally wrong.

In other cases, a practice of morally blaming agents for φ-ing, over time, generates desires not to φ that are more closely connected to the (presumed) moral wrongness of φ-ing. This can happen in various ways. One is by leading someone to apply a moral category that they already care about to something to which they did not previously apply it. For example, an agent might, as a result of being morally blamed, come to think that some activity they engage in is dishonest. Insofar as they were already motivated to avoid dishonesty, this could

generate a motivation in them to avoid the activity in question. In other cases, it may be that someone is led to see the moral relevance of a consideration that they previously thought was irrelevant. For example, perhaps someone might, through being morally blamed, come to see the importance of respecting others' privacy. (The distinction between this sort of case and the previous one may not always be sharp.)

I am now in a position to explain the claim that deontic moral concepts serve the Deliberative Reliability Function more fully, and explain what needs to be done to establish this claim. Some actions set back important interests, such as the interests we have in not being killed, hurt, or lied to. Insofar as we tend to judge such actions to be morally wrong, these judgments tend, over time, to bring it about that we are disposed to morally blame those who perform such actions without a moral excuse. This disposition, in turn, helps to bring about a continuous process of recruitment, by which people are led to be reliably motivated not to perform such actions (or to carry on being reliably motivated not to perform them). This, moreover, serves the further interests we have in being able to *rely on* not being treated in ways that set back our important interests. Given that using deontic moral concepts reliably produces these effects, and given these effects are valuable, deontic moral concepts serve the Deliberative Reliability Function.

I turn now to the tasks of showing that (1) using deontic moral concepts really does reliably produce these effects, and that (2) these effects really are valuable. With respect to the first task, I will not defend further the claim that our emotions are significantly, but imperfectly, responsive to our fittingness judgments. Instead, I will focus on the claim that a practice of morally blaming people for φ-ing tends, over time, to bring it about that people are reliably motivated not to φ.

This claim is empirically supported. In particular, work in psychology supports the claim that expressed blame (in particular, angry forms of expressed blame such as expressed resentment and indignation) often produces feelings of guilt in blamees (Baumeister et al. 1995; Parkinson 1999; Parkinson and Illingworth 2009). Moreover, and especially important for my purposes, expressed blame often produces feelings of guilt in blamees *even if blamees initially do not judge themselves to have acted morally wrongly* (Parkinson and Illingworth 2009). As we saw in the last chapter, guilt motivates agents to make amends through such things as confessions, apologies, offers of compensation, and, in many cases, commitments not to re-offend. Moreover, guilt is often effective in getting guilty agents not only to *commit* to not re-offending, but *actually* not to re-offend (Baumeister et al. 1995). So, there is empirical support for the claim that a practice of morally blaming people for φ-ing tends, over time, to bring about and sustain a situation in which people are reliably motivated not to φ (through inducing guilt-feelings in blamees).[7] Moreover, this is presumably not the only way in which a practice of morally blaming people for φ-ing can tend, over time, to produce reliable motivation not to φ. People can be motivated to avoid someone's blame even if it does not produce guilt in them – for instance, because they have self-interested reasons for wanting the potential blamer to be on good terms with them.

In order to show that using deontic moral concepts serves the Deliberative Reliability Function, we do not need to explain *why* a practice of morally blaming people for φ-ing tends, over time, to bring about and sustain a situation in which people are reliably motivated not to φ – we only need to show that a practice of morally blaming people for φ-ing in fact has this effect. Still, it will be worthwhile to explore this further question, in particular with respect to proleptic blame, since it is harder to see how moral blame can lead to appropriate motivation in agents who are not already disposed to recognise the force of the moral considerations to

[7] Further evidence for this claim may come from restorative justice programs. See Rossner (2019) for discussion.

which it adverts. I will begin with a proposed explanation of this by George Tsai (2017). Tsai's explanation, I will argue, captures part of what is going on, but also misses an important mechanism by which moral blame is effective.

Tsai's explanation of how blame tends to produce appropriate motivation in proleptic cases essentially involves two ideas: the first, which he draws from Williams (1995a: 41), is that proleptic blame typically relies for its efficacy on blamed agents desiring the good opinion and esteem of blamers; and the second is that agents can, over time and through habituation, come to be intrinsically motivated by moral considerations that previously motivated them only instrumentally (Tsai 2017: 260-265). Tsai's thought is that the desire for the good opinion and esteem of blamers furnishes blamees with an initial, instrumental motivation to act in accordance with the relevant moral considerations. Over time, as a result of habituation and a sustained practice of moral blame, blamees can subsequently come to care greatly and intrinsically about the relevant moral considerations.

It is surely correct that it usually takes time, habituation, and a sustained practice of moral blame to get people to care significantly and intrinsically about moral considerations they previously hardly cared about, or did not care about at all. Moreover, the suggestion that, for blame to get a grip in proleptic cases, the blamee must desire the good opinion and esteem of the blamer also seems plausible. Still, Tsai's explanation of the efficacy of proleptic blame leaves out something important. To see what this is, we will need to think more about the way in which morally blaming someone for φ-ing treats them as if they had significant normative reasons not to φ.

As we saw in Chapter 1, resentment and indignation motivate agents to make offenders feel guilty (for the right reasons) and make amends for their conduct (Dill and Darwall 2014: 46-54). Plausibly, an important element of making amends – in particular, of apologies – is

communicating to the addressee of the apology that you had significant normative reasons not to perform the relevant action. For example, suppose that someone professes to 'apologise' for being a bad spouse, but goes on to make it abundantly clear that they regard the standards they thereby transgressed as being normatively insignificant (that is, as not providing genuine normative reasons). Intuitively, this would not count as a sincere apology. Generalising, sincerely apologising for φ-ing requires communicating to the addressee of the apology that you had strong normative reasons not to φ.[8] Moreover, as the bad spouse example brings out, it is not enough to communicate that you had just *any* significant normative reasons not to φ. For example, to acknowledge only that you had strong normative reasons of self-interest not to φ would just add insult to injury. Rather, a sincere apologiser must communicate that the standards in virtue of violating which they are morally blameworthy themselves provided weighty normative reasons. In morally blaming someone, then, you are motivated to make them do something that requires them to communicate having had significant normative reasons to act otherwise. This social pressure towards apologising and making amends is what explains why moral blame treats the blamee as if they had strong normative reasons not to perform the relevant action.

This feature of other-directed moral blame, together with the fact that targets of moral blame are often prompted to feel guilty even when they initially judge themselves not to have acted culpably morally wrongly, helps to explain why moral blame can be effective in proleptic cases. Guilt-feelings, like emotions generally, significantly affect patterns of attention: in the case of guilt-feelings, attentional focus tends to be directed towards the object of your guilt and the violated standards that (appear to) make it blameworthy. Morally blaming others can

[8] At least, this seems to be true for a significant subset of apologies – what we might call serious apologies. These are apologies that express guilt-feelings, and which people are typically interested in receiving largely because they take these feelings to stand behind them. Serious apologies can be contrasted with more trivial apologies for minor social gaffes. In these cases, a simple 'I'm sorry' may be all that is required for an adequate apology, and it is not clear that such an apology conveys the message that you had any significant normative reasons to act otherwise.

insistently draw blamees' attention to the standards in virtue of violating which blamers take blamees to be blameworthy, and put social pressure on them to acknowledge the normative force of these standards. Getting moral blamees to attend to these standards in this socially pressurising way – particularly if there is an enduring, general practice of morally blaming people for culpably violating them – can, over time, get blamees to internalise these standards and see them as normatively important.

This is consistent with claiming that both habituation and blamees' desires for the esteem of blamers are important to explaining the efficacy of blame in proleptic cases. It seems plausible that moral blame will be most effective in drawing blamees' attention to the standards in virtue of violating which blamers take blamees to be blameworthy in cases where the esteem of the blamer is important to the blamee. And it may well be that moral considerations get a progressively stronger grip as blamees become habituated into acting on them. So, we need not reject the elements of Tsai's explanation of the efficacy of moral blame in proleptic cases. But appealing in addition to the claims that other-directed moral blame involves social pressure towards apologising and making amends, and that targets of moral blame are often prompted to feel guilty even when they initially judge themselves not to have acted culpably morally wrongly, gives us a fuller and deeper explanation of this phenomenon.

So far, I have argued that using deontic moral concepts reliably produces the effect of protecting important interests through encouraging reliable patterns of deliberation and motivation, such as the interests we have in not being killed, hurt, or lied to (and the further interests we have in being able to rely on not being treated in these ways). My next task is to show that this effect is valuable. Here, I can be brief. It is clearly valuable to use concepts that protect our important interests. Moreover, it is vitally important for the pursuit of most of the things we value that we are able to rely on our most basic interests not being threatened. As John Stuart Mill puts the point:

> …but security no human being can possibly do without; on it we depend for all our immunity from evil, and for the whole value of all and every good, beyond the passing moment; since nothing but the gratification of the instant could be of any worth to us, if we could be deprived of anything the next instant by whoever was momentarily stronger than ourselves. (Mill 1998 [1861]: V)

A shared practice of using deontic moral concepts is one way of bringing about security. It is not the only way: it can and should be supplemented by other means, such as an effective legal system. But deontic moral concepts make an important contribution to this end, and this is one reason why they are valuable.

This completes my positive argument for the claim that deontic moral concepts serve the Deliberative Reliability Function. I will now consider an objection to this way of arguing for the value of deontic moral concepts. It might be pointed out that the very same mechanisms by which deontic moral concepts are able to serve the Deliberative Reliability Function also enable them to have extremely pernicious effects if they are deployed in bad ways. For example, in a society in which homosexuality is typically judged to be morally wrong, using deontic moral concepts will tend to lead people to morally blame homosexuals, and this can be expected to lead homosexuals often to experience guilt-feelings about their homosexuality, against, or perhaps in accordance with, their better judgment. Or, to give a different example, if some members of a society accept a very stringent account of their positive moral obligations, this may lead them to experience persistent guilt-feelings about failing to live up to this ideal and inhibit them from developing well-rounded, meaningful lives involving commitments to projects outside of morality. This concern seems to be one strand in Williams's critique of what he calls 'the morality system' (1985: Chapter 10; 1995c: 205).

In response, it should be conceded that deontic moral concepts can have pernicious effects if they are deployed in bad ways. But this should not lead us to conclude that deontic moral concepts, as MB understands them, are not valuable concepts that are worth keeping. *Any* normative concept can have pernicious effects if it is deployed in bad ways. This is true of aretaic and evaluative concepts no less than deontic moral concepts. All normative concepts are such that their value is in some ways conditional on how they are deployed. Reflection on the pernicious effects of using deontic moral concepts in certain ways should not lead us to conclude that we would be better off without these concepts entirely, but rather to consider carefully what kinds of conduct are the proper objects of the distinctive forms of social pressure that use of deontic moral concepts brings with it.

3.2 The Respect Function

This section argues that deontic moral concepts, as understood by MB, serve a second valuable social function, which I call 'the Respect Function':

> *The Respect Function*: Use of deontic moral concepts reliably produces the effect of socially recognising the dignity of persons and their worthiness of respect.

My explanation of how deontic moral concepts serve this function runs via links between moral blame and making amends, and making amends and respect for persons. As with my discussion of the Deliberative Reliability Function, my discussion of the Respect Function is divided into two parts. I first (1) argue that a practice of using deontic moral concepts really does produce the relevant effect, and then (2) explain why this effect is valuable. I end with a brief discussion of how the argument of this section might be extended to respect for things other than persons, such as nature, and finally consider an objection to the arguments of both this section and the previous one stemming from scepticism about the value of anger.

Guilt, as we have seen, motivates agents to make amends through such things as apologies and offers of compensation. Resentment and indignation, in turn, motivate agents to get offenders to hold themselves accountable by feeling guilty (for the right reasons) and making amends. Moreover, we have seen that there are good reasons for thinking that resentment and indignation, alongside their expression in overt moral blame, are often effective in achieving this goal. A practice of using deontic moral concepts, then, can be expected to bring about that victims of wrongdoing fairly reliably receive apologies, compensation, and other forms of amends from wrongdoers. Moreover, insofar as making amends itself, or at least some of its elements, is judged to be morally required in the wake of unexcused moral wrongdoing, then we should expect this link to be even more robust (*via* the Deliberative Reliability Function).

As I stressed in the last section, to claim that a concept serves a valuable social function is not to claim that it *always* produces the relevant valuable effect; only that it tends reliably to do so under propitious circumstances. This bears re-emphasising, because it is obvious that not all moral wrongdoers are brought to feel guilty and make amends with victims through being morally blamed. But a practice of deploying deontic moral concepts, by engaging our dispositions towards moral blame, can be expected to bring about that victims of wrongdoing fairly reliably receive apologies, compensation, and other forms of amends from wrongdoers. Moreover, even in cases where guilt and amends are not forthcoming from wrongdoers, resentment and indignation, along with their expression in overt moral blame, can to some extent serve some of the same functions, such as affirming the worth of the victim.

Let me now turn to the link between making amends and respect for persons. It seems clear that making amends for φ-ing can be an expression of respect – in particular, but not exclusively, for the person *with whom* you make amends (Radzik 2009: 75-109; Jonker 2020: 30-32). But it is less clear *how* making amends expresses respect. As we will see, there are

various ways in which this happens. I will start by outlining these ways, before locating them with reference to a well-known distinction that Stephen Darwall has drawn between two kinds of respect: 'appraisal respect' and 'recognition respect' (1977, 2006a, 2012b).

We saw in the last section that sincerely apologising for φ-ing requires communicating to the addressee of the apology that you had strong normative reasons not to φ. Moreover, it is not enough to communicate that you had just *any* significant normative reasons not to φ. Rather, a sincere apologiser must communicate that the standards in virtue of violating which they are morally blameworthy themselves provided weighty normative reasons. Now, these standards typically make reference to certain salient facts about the victim – for example, that φ-ing significantly harmed them, or that they did not consent to the offender's φ-ing. This explains one way in which making amends for φ-ing expresses respect: it communicates to your victim (and others) that they are normatively significant, in the sense that certain salient facts about them, such as facts about their interests, are strong normative reasons to treat them in certain ways and not in others.

Another way in which making amends for φ-ing expresses respect is by involving submission to the authority of your victim to determine, within reasonable bounds, what counts as adequate amends (Walker 2006: 200-201). As Margaret Urban Walker puts it, wrongdoers '[relinquish] primary authority over… …the measure of satisfactory response' (2006: 200). The qualification 'within reasonable bounds' is important, because it is of course possible for victims to demand too much (or perhaps not enough) in the way of amends from moral wrongdoers. But within this range, victims have discretion to determine what will count as adequate amends. By 'adequate amends', I mean amends adequate to the offender and victim putting the wrong behind them, where this means, at a minimum, that it is no longer legitimate for the victim or others to demand further reparation from the victim or engage in any kind of

punitive response.[9] The point that making amends involves submitting to the authority of your victim to determine, within reasonable bounds, what counts as adequate amends, applies both to apologies and further reparative gestures, such as compensation. For an offender to quibble over precisely what counts as sufficient apology or compensation 'is likely to appear as denial or evasion and to add the proverbial insult to injury' (Walker 2006: 200).

Moreover, to make amends for φ-ing, it is not enough only to apologise, offer compensation, or attempt to re-form your character. Making amends also requires accepting the authority of your victim and, perhaps, members of the wider community, to demand these things of you (Darwall 2006a: 82-86; Walker 2006: 200). Suppose that someone apologises to their victim, offers compensation, and attempts to re-form their character, but regards all of these things only as pre-requisites of virtue, rather than things that their victim could legitimately demand of them. We might imagine that, in response to the demands of their victim for apology, compensation, and re-form, they reply, 'I'll do all of those things – but not because you expect them of me. I'll do them because I want to be a better person'. Such an agent would seem not to have adequately made amends. On the contrary, revealing that they see their attempts at amends in this light would be insulting. This points to another way in which making amends for φ-ing expresses respect: it expresses respect for the authority of your victim and, perhaps, members of the wider community, to demand such things as apology, compensation, and attempts at re-form from you.

Even in cases in which wrongdoers do not feel guilty and make amends, resentment and indignation, along with their expression in overt moral blame, can express respect for

[9] A further issue is whether forgiveness is warranted on the part of the victim. Forgiveness is usually thought to involve more than ceasing to demand further reparations from the victim or engage in punitive responses. For example, many hold that forgiveness requires abandoning or forswearing resentment (Murphy 1982; Hieronymi 2001; Griswold 2007: 38-43). See Hughes and Warmke (2022) for an overview of different theories concerning the nature of forgiveness.

victims of wrongdoing.[10] Resentment and indignation motivate agents to get offenders to feel guilty (for the right reasons) and make amends. Often, this is expressed by demands for apologies and other forms of amends. Even if these demands are not heeded by wrongdoers, the mere fact that they are made by victims and members of the wider community can itself convey the message that victims are worthy of respect, in light of the connections noted above between making amends and respect.

Darwall (1977, 2006a, 2012b) distinguishes 'recognition respect' from 'appraisal respect'. Whether someone has recognition respect for an agent or object, in the broadest sense, concerns how they give weight to features of that person or object in their practical reasoning and conduct. For example, a soldier might have recognition respect for their officer insofar as they give their orders significant positive weight as reasons in their practical deliberations and act accordingly. In contrast, to have appraisal respect for someone, in the broadest sense, is to appraise that person positively along an evaluative dimension, where the evaluative dimension in some way concerns their conduct or character. For example, someone might have appraisal respect for someone as a musician, where that involves a positive appraisal of such things as their musical abilities and dedication to their craft.

The ways in which making amends expresses respect that were distinguished above all concern recognition respect. To communicate to your victim that certain salient facts about them, such as facts about their interests, are strong normative reasons to treat them in certain ways and not in others is to communicate that they are worthy of recognition respect insofar as they possess these features. And submitting to the authority of your victim to determine, within reasonable bounds, what counts as adequate amends involves allowing your victim to shape

[10] The claim that indignation, and, especially, resentment can express respect for victims (in the case of resentment, self-respect) is a common theme in the literature on blame and forgiveness. See, e.g., Murphy (1982); Dillon (1997); Griswold (2007: 43-47); and Tierney (2019).

your practical reasoning, by treating their decision as to what counts as adequate amends as a reason for making amends in that way. Finally, accepting the authority of your victim and, perhaps, members of the wider community, to demand amends from you involves various deliberative and behavioural dispositions, in particular dispositions to acquiesce in, and not protest against, their legitimate demands.

The kinds of respect expressed by making amends are plausibly seen as central aspects of the recognition respect we owe to all persons as equals. Intuitively, it seems plausible that we owe amends not only to victims of unexcused wrongdoing who we happen to like, or who have high social status, but to all persons. Quite who counts as a person in this sense, and what the bases for personhood are, are controversial issues.[11] I will say more about this in a moment; for now, I will assume that it includes at least all cognitively able adult human beings (I will later claim that it includes all human beings *period*). The kinds of respect expressed by making amends do not exhaust the recognition respect we owe to all persons as equals. Having recognition respect for someone as a person involves not only dispositions towards making amends after unexcused moral wrongdoing, but also dispositions towards recognising and responding appropriately to reasons to treat them in certain ways and not in others in the first place. (Insofar as deontic moral concepts serve the Deliberative Reliability Function, they also contribute towards people being treated with this form of recognition respect). Moreover, recognition respect for persons may involve not only respect for the authority of victims to demand such things as apology, compensation, and attempts at re-form after unexcused moral wrongdoing, but also respect for the authority of people in general prospectively to demand compliance with moral requirements (Darwall 2006a).

[11] See Dillon (2022) for an overview of discussions concerning what respect for persons consists in and to whom it is owed.

Let me now bring together the ideas developed so far in this section to argue that a practice of using deontic moral concepts reliably produces the effect of socially recognising the dignity of persons and their worthiness of respect. A practice of deploying deontic moral concepts, by engaging our dispositions towards moral blame, can be expected to bring about that victims of wrongdoing often receive apologies, compensation, and other forms of amends from wrongdoers. Moreover, making amends for φ-ing can be an expression of respect – in particular, but not exclusively, for the person with whom you make amends – and this form of respect is plausibly seen as an aspect of the recognition respect we owe to all persons as equals. By expressing respect for others through our sincere attempts at making amends, we implicitly convey that they are worthy of respect. Dignity of persons and their worthiness of respect are mutually entailing notions: if someone has dignity as a person, they are worthy of respect as a person, and *vice versa* (Darwall 2006a: 120-121). So, a practice of using deontic moral concepts reliably produces the effect of socially recognising the dignity of persons and their worthiness of respect.[12] By 'recognising' something as X, I mean acknowledging, implicitly or explicitly, that something has or is X. A practice of using deontic moral concepts produces the effect of socially recognising the dignity of persons and their worthiness of respect by encouraging respectful forms of treatment – in particular, making amends with victims after unexcused moral wrongdoing. I use 'socially' to indicate that this often has a social dimension: in many cases, especially if the wrongdoing is relatively serious, knowledge that the offender has made amends will not be confined to the immediate recipients of amends.

I will now explain why the effect of socially recognising the dignity of persons and their worthiness of respect is valuable. First, this effect is in various ways instrumentally valuable. For example, victims who are treated respectfully by their offenders making amends

[12] This claim does not imply that our dignity as persons and worthiness of respect are somehow *constituted* by, or *resultant* from, there being a practice of using deontic moral concepts. This constitution claim is consistent with, but not entailed by, the claim that deontic moral concepts serve the Respect Function.

with them are more likely to be freed of some of the negative consequences of wrongdoing, such as fear, anger, and loss of trust, than victims who are not treated respectfully in this way. Moreover, *self*-respect, which can be fostered by social recognition of the dignity of persons and their worthiness of respect, is instrumentally valuable as well. Lacking self-respect can make people miserable. Moreover, it may be that self-respect is needed for certain valuable kinds of relationship, such as valuable forms of friendship. Someone entirely lacking in self-respect – who thought it did not matter whether others mistreated or abused them – would seem incapacitated for friendships in which friends view each other as equals. Finally, material compensation can of course be instrumentally valuable as well.

It also seems plausible that relating to persons as equals who have dignity and are worthy of respect is intrinsically valuable. Regarding people in this way, and appropriately responding to them as such, where this includes, importantly, regarding them as being owed amends if they are victims of unexcused moral wrongdoing (and making amends if you wrong them without an excuse), seems worth valuing for its own sake – and hence seems intrinsically valuable.[13] To see this, imagine that someone is culpably wronged, and that the offender, at first, does nothing to make amends for it. Eventually, however, they offer a sincere apology and make further attempts at amends, such as agreeing with their victim on (and carrying out) appropriate compensation. It seems that this would be worth valuing not only for its good effects, such as helping to free the victim from negative emotions like fear and anger, but also as a form of respectful treatment.[14]

[13] Although I am appealing to what is worth valuing for its own sake to support a claim about intrinsic value, I do not assume that intrinsic value is *analysable* in terms of what is worth valuing for its own sake. I only assume an equivalence between these things.

[14] The central idea here – that fitting responses (in this case, respect) to that which merits, or is worthy of them (in this case, dignity) are, or can be, intrinsically valuable – has precedent. See, for example, Thomas Hurka's (2001) recursive account of virtues as higher-order intrinsic goods that consist in responding fittingly to lower-order intrinsic goods and evils.

This concludes my positive argument for the claim that deontic moral concepts serve the Respect Function. There are good reasons for thinking that a practice of using deontic moral concepts reliably produces the effect of socially recognising the dignity of persons and their worthiness of respect. And there are good reasons for thinking that this effect is valuable. In the remainder of this section, I consider, first, whether the argument of this section might be extended to respect for things other than persons, such as the natural world, and, second, an objection to the arguments of both this section and the previous one flowing from scepticism about the value of anger.

It might be thought that my argument that deontic moral concepts serve the Respect Function, even if successful, at most shows that 'directed' deontic moral concepts are valuable – and, moreover, only insofar as these concepts are applied in cases in which victims are cognitively able adult human beings. A directed moral requirement is a moral requirement we owe *to* someone; correspondingly, when we fail to discharge such a requirement, we not only act morally wrongly, but wrong *them*. For example, when you morally wrongfully lie to someone, it is plausible to think that you wrong *them*. This goes hand in hand with the further thoughts that resentment on the part of the person you lied to is, *ceteris paribus*, fitting, and that they have special, perhaps unique, standing to forgive you.[15] Now, in my discussion of the connections between making amends and respect for persons, I focussed principally on how making amends expresses respect for *victims*. 'Victims', in one central sense of that term, are wronged parties. Moreover, many of the forms of respectful treatment I outlined depend on victims understanding communications, such as apologies, and also being capable of making demands and determining what counts as adequate amends.

[15] For further markers of directed moral requirements, see, e.g., Owens (2012: 44-67); Darwall (2013a: 20-39); Wallace (2019a: 5-11); and Jonker (2020: 3-16).

Given all of this, it might be thought that my argument that deontic moral concepts serve the Respect Function has a fairly limited application – applying only to directed deontic moral concepts, and only insofar as these concepts are applied in cases in which victims are cognitively able adult human beings. This would not show that the argument is of no interest – directed deontic moral concepts are a centrally important subclass of deontic moral concepts, and their application in cases in which victims are cognitively able adult human beings is a centrally important application of them. But we might have hoped for a further-reaching argument. I will now consider how my argument that deontic moral concepts serve the Respect Function might be extended. The discussion will of necessity be brief, but I will conclude, tentatively, that the argument can be significantly extended.

Let me start with the application of directed deontic moral concepts in cases where the victim is a human being, but not a cognitively able adult. For example, they may be a young child, or an adult suffering from severe cognitive impairments, such as those involved in severe Alzheimer's disease. We cannot make amends with such agents in the same way that we can make amends with cognitively able adults. We cannot offer apologies that they will (fully) understand; nor can we often sensibly make agreements with them concerning what forms of compensation or reparation would be appropriate. However, this does not mean that we cannot make amends with such agents at all. We can, for example, offer apologies to people who are saliently connected to the victims who can receive them on their behalf. Moreover, it is often possible for us to ameliorate the harms our wrongdoing caused through compensation, even if the victim is unable to deliberate with us concerning what forms appropriate compensation should take, or even recognise our reparative efforts *as* compensation.

Furthermore, these ways of making amends with victims who are not cognitively able adult human beings express respect for these people in much the same ways as making amends expresses respect in standard cases. Our apologies in these non-standard cases can

communicate that our victims are normatively significant, in the sense that certain salient facts about them, such as facts about their interests, are strong normative reasons to treat them in certain ways and not in others. Moreover, we can acknowledge the authority of people to demand amends on behalf of our victims. These forms of respectful treatment are in various ways instrumentally valuable – for example, in helping to free the families of victims from some of the negative consequences of wrongdoing, such as fear, anger, and loss of trust. Moreover, on the assumption that human beings who are not cognitively able adults are still worthy of respect as equal persons with dignity, a practice of making amends with such victims is plausibly intrinsically valuable. The relevant assumption is widely accepted (including by myself), even though it is famously difficult to explain the basis on which all human beings are worthy of respect as equal persons with dignity.[16]

Let me turn next to undirected moral requirements. Any example of such a requirement is bound to be controversial, but for the sake of argument I will focus on the moral requirement not to despoil areas of natural beauty. To adapt an example from T. M. Scanlon, imagine that someone filled the Grand Canyon with rubbish (1998: 219-220). Now, it may be that this would violate some directed moral duties – for instance, duties deriving from the interests that people have in visiting and enjoying the Grand Canyon. But we might also think that, apart from these interests, filling the Grand Canyon with rubbish would violate an undirected moral requirement not to despoil areas of natural beauty deriving from the impersonal value of such areas. This moral requirement is undirected, because it is not owed *to* these areas of natural beauty. We would not wrong the Grand Canyon by filling it with rubbish.

The form that appropriate amends take in cases where an undirected moral requirement of this kind is culpably violated is bound to be different from the form they take in cases where

[16] For relevant discussion, see, e.g., Scanlon (1998: 177-188); Arneson (2015); and Jaworska and Tannenbaum (2023).

directed moral requirements are culpably violated. For example, apologies addressed to areas of natural beauty, such as the Grand Canyon, would make little sense. Nor would it make much more sense to apologise to someone who could receive the apology on behalf of the Grand Canyon. In the case of victimised human beings who are not cognitively able adults, it is usually possible to identify someone who bears a significant relation to the victim and who could receive the apology on their behalf. Moreover, we can usually form an idea of how the victims might have been able to understand the apologies themselves, for instance if they were not suffering from Alzheimer's disease. Neither of these conditions seem to be met in cases where areas of natural beauty are despoiled.

However, this does not mean there are *no* ways in which we can appropriately make amends in cases where we have culpably violated an undirected moral requirement not to despoil areas of natural beauty. For example, we can try to restore these areas to their unspoiled states. Moreover, we can publicly acknowledge responsibility for wrongdoing – even if this acknowledgement, unlike standard apologies, is not addressed to anyone in particular. Finally, it is possible for others legitimately to demand these things of us, and for us to accept their demands as legitimate. In all of these ways, we can appropriately make amends when we culpably violate an undirected moral requirement of this kind.

These ways of making amends express respect for areas of natural beauty. In acknowledging responsibility for wrongdoing, and attempting to undo the damage we have caused, we acknowledge that we have strong reasons to protect and preserve areas of natural beauty, and not despoil them. Moreover, accepting that it is, or would be, legitimate for others to demand these responses of us also expresses respect. Insofar as areas of natural beauty are worthy of respect, a practice of making amends in these ways is plausibly intrinsically valuable. Moreover, it is in various ways instrumentally valuable – for example, in bringing it about that people are able to enjoy these areas again. So, it looks as though the argument that deontic

moral concepts serve the Respect Function might plausibly be extended beyond respect for persons to respect for other things, such as natural beauty.

Let me now move on to a different issue. Resentment and indignation, as understood here, are forms of anger. There is a long tradition of scepticism about the value of anger, and it is worth considering my argument that deontic moral concepts are valuable in virtue of serving the Deliberative Reliability Function and the Respect Function in light of this tradition. I will focus in particular on Martha Nussbaum's (2016) recent and forceful defence of an anti-anger position.[17]

Nussbaum argues that anger characteristically involves a *payback wish*: a wish for things to go badly, in some way, for the target of the anger, in a way that is imagined somehow to constitute revenge or payback for their offense (21-27). She distinguishes two sources of this payback wish (24-27). The first is in magical thinking that the suffering of the offender somehow balances or undoes the harm that was initially done by the offense. The second is in the non-magical, often correct thought that inflicting suffering on the offender helps to restore the relative social status of the wronged party by humiliating the offender and down-grading their relative social status.

Nussbaum concedes that ordinary anger, involving the payback wish, has some, albeit limited, social utility (37-40). First, it is valuable as a signal, both to oneself and to others, that a wrong has been done. Second, it is valuable as a source of motivation in combating injustice. Third and finally, it is valuable in deterring others from wrongful acts. While anger is valuable in all these ways, Nussbaum thinks that its value is limited because it includes the payback wish, which is unhelpful and destructive in both of its forms. Inflicting harm on wrongdoers

[17] See Pettigrove and Tanaka (2014: 272-277) for an overview of some anti-anger positions in the history of philosophical and religious thought. Nussbaum's defence of an anti-anger position is representative of this tradition in its focus on the claim that anger involves a desire for payback or revenge.

cannot balance or undo their wrongful actions (24-25). And pre-occupation with relative social status should be discouraged on the grounds that there are more important values deserving of our attention (28-29). Both forms of the payback wish distract us from constructive attempts to ensure that similar wrongs do not happen on future occasions. Nussbaum goes on to identify a borderline case of anger she calls 'Transition Anger', which she argues is much more valuable (35-37). It is a borderline case because it does not involve the payback wish. Instead of seeking payback, it focuses on promoting social welfare.

As I explained in Chapter 2, I use 'resentment' and 'indignation' as terms of art for forms of anger that are distinctive in virtue of their motivational goals. Both aim at making offenders hold themselves accountable by feeling guilty (for the right reasons) and making amends. The goal of resentment is to make the offender feel guilty about what they have done to *you* and make amends with *you*. The goal of indignation, on the other hand, is to make the offender feel guilty about what they have done with respect to some other person(s) or impersonal value and make amends in the appropriate way. Dill and Darwall (2014) review extensive empirical evidence for the claim that these emotions are psychologically real and the typical driving force behind condemnatory behaviour.

Resentment and indignation do not involve the payback wish. Nor are they purely forward-looking like Transition Anger, concerned only with promoting social welfare.[18] They are both backward-looking and forward-looking: backward-looking insofar as they aim to get offenders to hold themselves accountable for a past deed, and forward-looking insofar as they aim to get them to make amends for it. We can agree with Nussbaum in criticising the forms of anger that she focusses on, while arguing that instead of banishing ordinary anger entirely we would be better off trying to re-direct and cultivate it into resentment and indignation, which

[18] The point that the forms of anger that Nussbaum distinguishes are not exhaustive is also made by Shoemaker (2018: 72-75); Srinivasan (2018: 7-8); and Wallace (2019b: 541-542).

are socially useful (I explain in a moment why Transition Anger cannot do the work of resentment and indignation).

Against these claims, it might be argued that resentment and indignation do involve the payback wish in both of the forms that Nussbaum distinguishes, albeit in a concealed, ritualised way. The practice of making amends through confession, apology, and reparation might be claimed to be humiliating and abasing (cf. Nussbaum 2016: 10-13, 57-75). So, perhaps resentment and indignation, like ordinary anger generally, aim at balancing the harm that was initially done by the offence and restoring the relative social status of the victim, through the humiliating rituals of confession, apology, and reparation (cf. 74).

However, making amends need not involve humiliation and abasement. For example, apologising need not involve characterising oneself in self-abasing terms, such as being worthless or disgusting. On the contrary, in typical cases making amends is primarily *victim-focussed*: it involves communicating that certain salient facts about your victim, such as facts about their interests, are strong reasons to treat them in certain ways and not in others; and it involves finding ways of compensating your victim for harm or damage. Of course, making amends is concerned with the offender insofar as it involves them admitting fault. But this need not involve self-abasement. For example, it is entirely possible for an offender to be properly repentant, and sincerely committed to making amends, while rejecting certain demands for reparation or punishment as cruel, excessive, and incompatible with the respect they deserve as an equal human being. Moreover, there is empirical evidence that what resentful and indignant agents typically seek from amends is not the abasement and humiliation of offenders. Schmitt et al. (2004) examined how effective different components of apologies are in reducing angry blame. They found that apologies that included admission of fault, expression of guilt/remorse, and, especially, admission of damage or harm and offer of compensation were most effective in reducing angry blame. None of the language used in these apologies was self-

abasing. The apologisers did not, for example, characterise themselves as worthless, lowly, or disgusting. A reasonable interpretation is that what the angry blamers wanted most was not abasement but acknowledgement: they wanted the offender to understand the damage they had caused, understand why it was important, and seek to repair the damage through material compensation.[19]

Resentment and indignation, then, do not involve the payback wish. Moreover, they are able to have their positive effects precisely because, unlike what Nussbaum calls 'Transition Anger', they are not purely forward-looking (cf. Strawson 1962/2003: 93). As we saw in the last section, it is because resentment and indignation focus offenders' attention on the standards they violated and put social pressure on them to accept the normative significance of these standards that they can be effective in changing offenders' motivation and behaviour in the future. Moreover, it is because they help to bring about that victims receive amends for past wrongdoing that they help to provide social recognition of the dignity of persons and their worthiness of respect.

3.3 Conclusion

This chapter argued that MB meets the prescriptive desideratum on conceptual analyses explained and defended in the Introduction. Deontic moral concepts, as analysed by MB, are extremely valuable concepts. This is because they serve two important social functions: the Deliberative Reliability Function and the Respect Function.

[19] Admittedly, the evidence for this interpretation would be stronger if the experimenters had included a self-abasement component and studied its effects. However, the fact that angry blame tended to be significantly reduced by admission of fault, expression of guilt/remorse, admission of damage or harm, and offer of compensation in the absence of self-abasement does tell against the claim that resentful and indignant agents want offenders to feel guilty and make amends primarily as a means of abasement.

Chapter 4

Extensional Objections

MB holds:

> *Moral Wrongness as Moral Blameworthiness (MB)*: It is morally wrong for an agent to φ if and only if (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, where 'φ' stands for an object of deontic moral assessment.

An important family of objections to MB targets its extensional adequacy. According to these objections, MB cannot provide a plausible conceptual and metaphysical analysis of moral wrongness, because the left-hand side of the bi-conditional stated above does not necessarily coincide with its right-hand side. In this chapter, I address extensional challenges to MB and consider what changes, if any, need to be made to MB to meet them. I argue that, with the exception of one extensional challenge, no changes need to be made to MB. Moreover, defenders of MB can meet these challenges while remaining largely neutral on the first-order debates at issue.

Section 1 motivates a neutrality desideratum for conceptual analyses in meta-ethics. Section 2 identifies two corollaries of MB. Sections 3-6 discuss and rebut four extensional challenges to MB. These challenges arise from the Objectivism/Subjectivism debate, the debate surrounding the possibility of suberogation, the debate surrounding the relevance of motivating reasons to moral wrongness, and the consequentialist tradition. I argue that, with the exception of challenges arising from the Objectivism/Subjectivism debate, no changes need to be made to MB to meet them successfully.

4.1 Neutrality

MB is an analysis of the concept and property of (all-things-considered) moral wrongness.[1] Analyses of normative concepts and properties can be contrasted with 'first-order normative views', where a first-order normative view is a claim about which things have (or do not have) certain normative properties and, perhaps, what explains why they have (or do not have) them. First-order normative views may link having certain non-normative properties with having certain normative properties (as in, 'It is morally wrong to torture cats for fun'), or they may link having certain normative properties with having certain further normative properties (as in, 'If an act fails to maximise value, it is morally wrong'). Moreover, first-order normative views, on this expansive understanding of them, may include various logical constants, such as conjunction, disjunction, and negation.

Any analysis of a normative concept and/or property has *some* first-order implications. This is because any such analysis implies that anything that satisfies its right-hand side has this property, and that nothing that does not satisfy its right-hand side has this property.[2] Given this, analyses of normative concepts and/or properties – if such analyses are to be possible at all – cannot aim for complete neutrality with respect to first-order normative views. Moreover, there may be some first-order normative views with respect to which normative analyses *should not* remain neutral. Some philosophers argue that certain first-order normative views are conceptual truths (Foot 1958; Cuneo and Shafer-Landau 2014). For example, it might be claimed that it is a conceptual truth that it is *pro tanto* morally wrong to humiliate others simply

[1] Strictly speaking, the metaphysical side of MB is an analysis of the moral wrongness *relation*. However, for the sake of readability, I will talk about properties.

[2] This does not mean, however, that error theories are ruled out. This is because the first-order views entailed by analyses of normative concepts and/or properties are *conditional claims*: *if* the right-hand side of the analysis is satisfied, *then* its left-hand side is; *if* the right-hand side of the analysis is not satisfied, *then* its left-hand side is not satisfied either. An error theorist could accept these claims while denying that the right-hand side of the analysis in question is ever satisfied.

for pleasure (Cuneo and Shafer-Landau 2014: 405). If there are conceptual truths like this, then a successful analysis of the relevant normative concept needs to explain them.[3]

Nonetheless, defenders of analyses of normative *concepts*, at least, should typically be very wary of taking on first-order commitments. (I leave it open whether analyses of normative *properties* are subject to as strong a neutrality desideratum). To see why this is so, we will need to revisit the account of conceptual analysis defended in the Introduction. I argued that a fruitful style of conceptual analysis is that which aims to jointly satisfy a descriptive desideratum and a prescriptive desideratum. The descriptive desideratum is that the analyses in question should be broadly in line with the concepts we in fact use. More carefully, conceptual analyses should for the most part respect the platitudes surrounding the relevant concepts, where 'platitudes' are statements of the judgmental and inferential dispositions that are typically had by those who possess the relevant concepts. The prescriptive desideratum on conceptual analyses is that the analyses should help to explain why the concepts in question are valuable concepts that are worth keeping.

It is because of the descriptive desideratum on successful conceptual analyses that they are subject to a strong neutrality desideratum. As we saw in Chapter 2, apparently denying a platitude associated with a concept, or making a judgment or inference that directly contradicts such a platitude, is evidence that the agent does not possess this concept, or at least is not employing it.[4] Moreover, appearing to deny a platitude or make a directly contradictory judgment or inference tends to produce a feeling of bafflement among those who possess the relevant concept. A couple of examples will illustrate this. A relatively uncontroversial

[3] In the case of MB, then, if there are such conceptual truths concerning what is morally wrong, it will need to be shown that these conceptual truths are based on conceptual truths concerning what would be morally blameworthy if done without a moral excuse (cf. Gibbard 1992: 211-213). I will not discuss this issue further in this thesis.

[4] The qualifier 'apparently' is needed because if the seeming denial of a platitude (or the seeming making of a directly contradictory judgment or inference) rightly leads us to conclude that the agent in question does not in fact possess the relevant concept, then we cannot correctly interpret them as making judgments containing this concept.

example of a platitude associated with our normative concepts is that there can be no normative difference without a non-normative difference: two actions that are exactly the same in non-normative respects must be the same in normative respects (Jackson and Pettit 1995: 21). If someone appears to make a judgment that directly contradicts this platitude – for example, by seemingly judging that action A was morally wrong and that action B was not morally wrong, while claiming that A and B were identical in non-normative respects – this counts as evidence that they do not possess the concept at issue (namely, MORAL WRONGNESS) and will tend to produce a feeling of bafflement among those who possess this concept. To give a non-ethical example, if someone apparently judges that a figure is a four-sided triangle, this counts as evidence that they do not possess the concept TRIANGLE, or at least are not employing it, and those who possess this concept will tend to feel baffled.

Conceptual analyses, I claimed, should aim for the most part to respect the platitudes associated with the target concepts. This means that a defender of an analysis of a normative concept with a certain set of first-order implications, F, is usually committed to claiming that F can be established solely on the basis of the relevant set of platitudes.[5] This is a heavy dialectical burden. It typically commits defenders of analyses of normative concepts not only to claiming that first-order views that directly conflict with their analyses are false. More strongly, it typically commits them to claiming that they can be shown to be false solely by reference to a set of judgments and inferences such that apparently contradicting any of them is evidence that the agent does not possess the normative concept in question. Given that first-order implications carry this heavy dialectical burden, defenders of analyses of normative concepts should typically be very wary of taking on such implications. Indeed, in the case of highly contentious first-order normative debates, where there are various fairly plausible first-

[5] 'Usually', because the prescriptive desideratum allows conceptual analyses to depart *somewhat* from the platitudes associated with the relevant concepts.

order views that have been subjected to much critical scrutiny, it would be a very significant mark against an analysis of a normative concept if it purported to resolve these debates on its own. It would be very surprising if such debates could be settled solely by appealing to the platitudes associated with the relevant concepts.[6]

4.2 Two Corollaries of MB

MB has two corollaries. They are:

> *Corollary 1*: If it is morally wrong for an agent to φ, then φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them.

And:

> *Corollary 2*: If φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, it is morally wrong for the agent to φ.

Most of the discussion in Sections 3-6 will focus on Corollary 2. In the interests of readability, I will discuss a simplified version of Corollary 2:

[6] It might be wondered how the relatively modest approach to conceptual analysis outlined in this section compares with Frank Jackson and Philip Pettit's well-known defences of conceptual analysis in meta-ethics (Jackson and Pettit 1995; Jackson 1998). Jackson and Pettit aim to defend a naturalistic reduction of moral properties through a holistic analysis of our network of moral terms. Such an analysis would need to have extensive first-order implications to facilitate such a reduction. This may seem to put Jackson and Pettit's approach to conceptual analysis in stark contrast to my own. However, it is possible that the difference between these projects is not a substantive difference, but only a reflection of differences in how we use the label 'conceptual analysis'. Jackson in particular sometimes characterises the inputs to his conceptual analyses in a way that includes much more than what I am calling 'platitudes'. At one point, he characterises the basic input to his conceptual analyses as 'what we find intuitively plausible' (1998: 135; but cf. also 130). This includes much more than what is covered by the platitudes associated with our moral concepts, since there are claims that many people find intuitively plausible but which are not such that failing to share these intuitions is evidence that the agent lacks moral concepts.

Corollary 2-S: An agent is morally blameworthy for φ-ing only if φ-ing is morally wrong.

Discussing this simplified version of Corollary 2 will not change anything of philosophical significance. Since Corollary 2-S will be the main focus of discussion, I will give it a name for the sake of readability: 'Blameworthy only if Wrong'.

4.3 Objective and Subjective Moral Wrongness

The first objection to the extensional adequacy of MB that I will consider concerns the relevance of an agent's beliefs and evidence to, respectively, moral wrongness and moral blameworthiness. More specifically, the concern is that the beliefs that an agent has when they φ and the evidence that is available to them are irrelevant to whether φ-ing is morally wrong, but relevant to whether they are morally blameworthy for φ-ing. The view that the moral wrongness of A's φ-ing does not depend on A's beliefs or evidence but only on A's objective circumstances is commonly known as 'Objectivism'. The objection, then, is that MB is false because, given Objectivism, moral wrongness and moral blameworthiness can come apart in ways that are inconsistent with MB in cases where agents act in conditions of error, ignorance, or uncertainty.

This extensional challenge to MB threatens Blameworthy only if Wrong rather than Corollary 1. If φ-ing is objectively morally wrong, but the agent is morally blameless for φ-ing because of error, ignorance, or uncertainty, this will be because they fail to meet some relevant epistemic condition on moral responsibility. But then such cases will not be counter-examples to Corollary 1, because the agent has a moral excuse. So, the objection to the extensional adequacy of MB from Objectivism calls only Blameworthy only if Wrong into question.

I will argue that the objection from Objectivism ultimately does not threaten MB. Given a minor modification, MB can be made compatible with Objectivism. First, however, it will help to have a couple of examples to illustrate the concern:

> *Murderous Husband*: Alfred's wife is dying and he wishes to hasten her death. He gives her a certain substance in the reasonable belief that it is poison (he also believes that giving it to her is morally wrong). Unbeknownst to him, it is the only existing cure for his wife's ailment. (Thomson 1991: 293)

> *Inadvertent Rescue*: Dan, a small child, is in danger. Paul does not realise that Dan is in danger or have any reason to believe that he is. Paul picks Dan up and runs off with him anyway, thereby rescuing him from danger. Paul does this because he wants to upset Dan, even though he believes that upsetting him is morally wrong. (Zimmerman 1997: 236)

Objectivists claim that Alfred does not act morally wrongly in giving his wife the substance and that Paul does not act morally wrongly in picking Dan up and running off with him. On the contrary, since these actions save Alfred's wife and Dan they are morally required. But it might nonetheless be thought that Alfred is morally blameworthy for giving his wife the substance and that Paul is morally blameworthy for picking Dan up and running off with him.[7] After all, both agents performed these actions while reasonably believing that they were thereby acting morally wrongly (and, we can add, they performed these actions freely). To accept both of these claims, however, is to give up on Blameworthy only if Wrong: an agent is morally blameworthy for φ-ing only if φ-ing is morally wrong.

[7] Against this, it might be argued that Alfred and Paul are not morally blameworthy for the actions they perform, but only for the bad practical reasoning that led them to perform these actions (cf. Portmore 2021: 56-57). But as we will see in a moment, even if Alfred and Paul are morally blameworthy for the actions they perform, a simple modification to MB allows it to avoid these counterexamples.

One response to this extensional challenge is to reject Objectivism. As we have seen, Objectivism holds that the moral wrongness of A's φ-ing does not depend on A's beliefs or evidence but only on A's objective circumstances. This contrasts with 'Subjectivism', according to which the moral wrongness of A's φ-ing depends on A's beliefs or evidence.[8] A Subjectivist would claim that Alfred and Paul act morally wrongly in *Murderous Husband* and *Inadvertent Rescue* because of their beliefs or evidence. Many defenders of MB reject Objectivism in favour of some version of Subjectivism (Gibbard 1990: 42-43; Skorupski 2010: 297). However, given the neutrality desideratum defended in Section 1, it would be greatly preferable to show that MB is compatible with either Objectivism or its denial.

A minor modification to MB renders it compatible with Objectivism. The modification adds a clause to MB concerning the epistemic situation of the agent:

> *MB-Objectivism*: It is morally wrong for an agent to φ if and only if (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse *and in full knowledge of the relevant features of the situation*, they would be morally blameworthy for violating them.

MB-Objectivism gives the result that Alfred does not act morally wrongly in giving his wife the substance and that Paul does not act morally wrongly in picking Dan up and running off with him. If Alfred gave his wife the substance in the knowledge that it is the only existing cure for his wife's ailment, he would not be morally blameworthy for doing so. And if Paul picked Dan up and ran off with him in the knowledge that doing so would rescue him from danger, he would not be morally blameworthy for doing this. Hence, MB-Objectivism is immune to these purported counter-examples to MB.

[8] Subjectivists differ as to whether the relevant beliefs or evidence concern only what outcomes will result from the actions available to the agent, or also the values of these outcomes. For discussion, see Zimmerman (2008: 37-42). I shall abstract from this complexity in what follows.

Just as it is possible to render MB compatible with Objectivism by introducing this minor modification, it is also possible to render it compatible with different versions of Subjectivism by adding clauses concerning the epistemic situation of the agent. It is sometimes argued that moral blameworthiness depends on the beliefs the agent in fact had, but moral wrongness depends on the beliefs that would have been evidentially warranted in their situation (Skorupski 2010: 295-297). Here is an example to illustrate this:

> *Bad Doctor*: Doctor gives her patient a substance. She believes that this substance will kill the patient, but her evidence indicates that it will cure them. The substance cures the patient. (Skorupski 2010: 297)

Some Subjectivists – in line with recent tradition, I will call them Prospectivists – argue that Doctor does not act morally wrongly in giving her patient the substance. On the contrary, since her evidence indicates that the substance will cure them, giving them the substance is morally required. At the same time, it might be thought that, given her belief that it will kill them, Doctor is morally blameworthy for giving her patient the substance. But to accept both of these claims is to reject Blameworthy only if Wrong: an agent is morally blameworthy for φ-ing only if φ-ing is morally wrong.

MB can be rendered compatible with Prospectivism through a small modification. The modification is as follows:

> MB-Prospectivism: It is morally wrong for an agent to φ if and only if (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse *and from the beliefs that are evidentially warranted in their situation*, they would be morally blameworthy for violating them. (cf. Skorupski 2010: 292)

MB-Prospectivism yields the result that Doctor does not act morally wrongly in giving her patient the substance. If Doctor gave her patient the substance in the evidentially supported

belief that it will cure them, she would not be morally blameworthy for doing so. Hence, MB-Prospectivism is not vulnerable to this putative counter-example to MB.

In summary, whatever the outcome of the debate between Objectivists and Subjectivists about moral wrongness, it will not lead to insuperable challenges to the extensional adequacy of MB. MB can be modified to be compatible with the main views in this debate by adding clauses concerning the epistemic situation of the agent.

4.4 Suberogation

A further objection to Blameworthy only if Wrong comes from suberogation. An action is suberogatory just in case it is morally permissible but (all-things-considered) morally bad (Driver 1992). Since suberogatory actions are morally bad, we might think that agents are morally blameworthy for performing them without a moral excuse. But to accept this would be to reject Blameworthy only if Wrong: an agent is morally blameworthy for φ-ing only if φ-ing is morally wrong.

The claim that there are suberogatory actions presupposes a distinction between moral and non-moral badness. Everyone should agree that there are actions that are morally permissible but bad *in some respect*. For example, it is usually morally permissible for me to tie my shoelaces with a bad knot, but this is not morally bad. The interest of the category of suberogation is that the actions in question are meant to be distinctively morally bad without being morally wrong. It is not entirely clear how this allegedly distinctive category of moral badness is to be understood, but I will leave this worry about suberogation to one side and suppose that an adequate account of it can be given.

It will help to start by looking at some putative examples of suberogation. Consider the following two cases, both of which are taken from Julia Driver's (1992) influential discussion of this topic:

> *Kidney Transplant*: Roger and Bob are brothers. Bob needs a kidney transplant and Roger is the only available donor. Roger refuses to donate his kidney to Bob even though he knows his brother will die without it. (Driver 1992: 287)

> *Train Seat*: Beatrice, Kate, and Owen have boarded a train and are looking for seats. It is obvious that Kate and Owen are a couple and would like to sit together. There are only two free seats next to each other in the carriage and Beatrice sits in one of them, fully aware that this means Kate and Owen will have to sit apart. (Driver 1992: 286-287)

In both of these cases, we might think that the agents act morally badly but morally permissibly. Or, at least, we might think that this is true of some versions of these cases, depending on how further details are filled in. Since these actions are morally bad, there must be some critical reactions that are fitting to agents who perform them without an excuse. But if moral blame is among these reactions, then we must reject Blameworthy only if Wrong (an agent is morally blameworthy for φ-ing only if φ-ing is morally wrong).

Given the neutrality desideratum defended in Section 1, it will not do for defenders of MB to respond to these worries by denying that there are any suberogatory actions. Rather, defenders of MB need to uphold Blameworthy only if Wrong while leaving room for the possibility of suberogation. I will argue that this can be done by exploiting the pluralist view of blame defended in Chapter 2, according to which there are various kinds of critical reactions that can constitute blame. I argued that the kinds of blame with which MB is concerned are guilt, resentment, and indignation, and I argued further that there are various other emotional

and non-emotional critical reactions that can constitute blame besides this trio. Moreover, I argued that indignation is the central moral blaming emotion. There are actions in response to which guilt and resentment would be fitting, but not indignation, and such actions are plausibly not morally wrong (I pointed to violations of certain standards of good friendship as examples of this). Finally, we saw in the Introduction that defenders of MB should appeal not simply to the fittingness of indignation, but to whether indignation would be fitting *from an impartial standpoint*. Let me call indignation that is felt from an impartial standpoint 'impartial indignation'. Now, it is open to defenders of MB to allow that there are suberogatory actions as long as they deny that impartial indignation is among the critical reactions that would be fitting in response to agents who perform these actions without an excuse.

Defenders of MB can respond to any putative example of suberogation as follows. Any putative example of suberogation is either a genuine example of suberogation or not. If it is not a genuine example of suberogation, then it does not threaten Blameworthy only if Wrong. If it is a genuine example of suberogation, it only threatens Blameworthy only if Wrong if impartial indignation is among the critical reactions that would be fitting in response to it. But it is always open to defenders of MB to point to other kinds of fitting blame to capture the moral badness of these actions.[9] Now, it is *prima face* highly plausible that fitting impartial indignation entails moral wrongness. Given this, we should be confident that other kinds of blame besides impartial indignation exhaust the kinds of blame that would be fitting in response to suberogatory actions (if there are any). This is because it will always be more plausible that fitting impartial indignation entails moral wrongness than it is that the example in question is a genuine case of suberogation. Let me explain and illustrate this argumentative strategy by applying it to the two putative examples of suberogation given above.

[9] Cf. Calhoun (2015: 109): '[agents who perform suberogatory actions are] open… …to the charge of being petty, mean-spirited, contemptible, disappointing, irritating, and a poor excuse for a moral agent'.

I will start with *Kidney Transplant*. A notable feature of this case is that Roger and Bob are brothers. Let us suppose that they get along well and are close, and belong to a culture that places significance on fraternal relationships. It seems to be an assumption of the example that if Roger and Bob were not brothers but strangers then Roger's action would be morally neutral. If *Kidney Transplant* is to be a successful illustration of suberogation, then close personal relationships such as brotherhood must have *some* moral significance, but not *enough* moral significance to generate a moral requirement in this and similar cases. Rather, Roger's relationship with Bob must only make it morally bad for him to refuse to donate his kidney without making this morally wrong.

It may be that *Kidney Transplant* and similar cases involving failings within close relationships are not genuine cases of suberogation. Perhaps these actions are either morally wrong or morally neutral, without there being any middle territory occupied by suberogation. If, on the other hand, *Kidney Transplant* and similar cases are genuine examples of suberogation, then they only undermine Blameworthy only if Wrong if the actions in question merit impartial indignation. However, it is open to defenders of MB to point to other kinds of blame that could reflect the moral badness of these actions, and it is more plausible that fitting impartial indignation entails moral wrongness than it is that *Kidney Transplant* and similar cases are genuine examples of suberogation.

A kind of blame that is especially important in the context of close personal relationships is the kind that T. M. Scanlon emphasises: making adjustments to your relationships with others, such as withdrawing trust and being less willing to help them in their projects, in light of how their actions reveal attitudes that impair the relationships you can have with them (2008). Scanlon illustrates this kind of blame with an example in which you learn that a 'friend' made a cruel joke about you at a party (129-131). For you to blame your 'friend', on Scanlon's view, is to adjust your attitudes towards them in ways that reflect how their joke

and the attitudes towards you it reveals impair your relationship with them. For instance, you might no longer value spending time with them, and you might no longer intend to confide in them or encourage them to do the same with you (Ibid.).

This kind of blame is surely an important element of our responses to failings within close personal relationships. If in the context of such relationships there are actions that are morally bad but not morally wrong, a defender of MB might argue that it is only Scanlon-style blame (and perhaps also guilt and resentment), but not impartial indignation, that is fitting. This would allow a defender of MB to capture the moral badness of these actions while denying that they are counter-examples to Blameworthy only if Wrong. It might be urged in response that there are cases in which *both* Scanlon-style blame *and* all of guilt, resentment, and impartial indignation are fitting, but the actions in question are nonetheless merely morally bad. But this would be dialectically ineffective. It is more plausible that fitting impartial indignation entails moral wrongness than it is that there are any such actions.

To see this, let us return to *Kidney Transplant*. Assume for the sake of argument that Roger acts morally badly but not morally wrongly in refusing to donate his kidney to Bob. A defender of MB might aim to capture the moral badness of his action by appealing to Scanlon-style blame. It certainly seems appropriate for Bob, in the weeks prior to his death, to reassess his relationship with Roger and adjust his attitudes towards him accordingly. For example, he might no longer be especially pleased when things go well for him. And perhaps it would also be fitting for Bob to resent Roger and for Roger to feel guilty. So, a defender of MB could accommodate the claim that there is something morally bad about Roger's action by appealing to some or all of these forms of blame, without claiming that Roger's action merits impartial indignation. But now suppose that someone were to insist that Roger is an apt target of Scanlon-style blame, resentment, guilt, *and in addition* impartial indignation. This claim would be dialectically ineffective. A defender of MB can already make room for the claim that Roger's

action is morally bad, and it is more plausible that fitting indignation entails moral wrongness than it is that Roger's action is not morally wrong.

Next, let me apply the same general argumentative strategy to *Train Seat*. The description of this case is relatively sparse and it might be thought that there are some ways of filling in the details such that Beatrice's action is clearly morally wrong. Hallie Liberto imagines a variant of the case in which it is the couple's last train ride before one of them goes to fight in a war (2012: 400). Here, we might think that it is sufficiently important to the couple to sit together that it is morally wrong for Beatrice to force them to sit apart. But suppose that there is no special reason for the couple to sit together; they simply want to sit together. And suppose further that there is no special reason for Beatrice to sit in one of the two available adjacent seats. If we fill in the details this way, it might seem tempting to claim that Beatrice's action is not morally wrong, but nonetheless morally bad in virtue of being inconsiderate. But if Beatrice would nevertheless be a fitting target of impartial indignation, then Blameworthy only if Wrong is in trouble.

However, as with *Kidney Transplant*, it is more plausible that fitting impartial indignation entails moral wrongness than it is that *Train Seat* provides a genuine example of suberogation. The question of how sharp a line should be drawn between manners and morals is contested (Buss 1999). For example, Sarah Buss argues that courtesy forms part of the moral requirement to treat others with respect (1999: 796). If this is right, then Beatrice acts morally wrongly in *Train Seat* by violating this moral requirement. It is hard to see how we could settle the question of the relation between manners and morals in a principled way without considering the similarities and differences between paradigmatic examples of moral requirements and paradigmatic examples of requirements of manners, and seeing whether the similarities are striking enough to count manners as part of morals (cf. Buss 1999: 803-804). But a close link with fitting impartial indignation is a central mark of paradigmatic examples

of moral requirements. Insofar as inconsiderate actions such as Beatrice's in *Train Seat* merit impartial indignation, in contrast to other negative reactions such as annoyance or irritation (or even guilt and resentment), then this would be strong evidence that inconsiderate actions are also morally wrong.

To summarise: defenders of MB should not be worried by suberogation. Any putative example of suberogation is either a genuine example of suberogation or not. If not, it is not a counter-example to Blameworthy only if Wrong. If it is a genuine case of suberogation, it only makes trouble for Blameworthy only if Wrong if impartial indignation is among the reactions that would be fitting in response to it. But it is always open to defenders of MB to point to other kinds of fitting blame to capture the moral badness of these actions. Given the strong *prima facie* plausibility of the claim that fitting impartial indignation entails moral wrongness, we should be confident that this strategy will succeed generally. It will always be more plausible that fitting impartial indignation entails moral wrongness than it is that any putative example of suberogation is a genuine example of suberogation.

4.5 Motivating Reasons

A third objection to the extensional adequacy of MB concerns the relevance of motivating reasons to moral wrongness and moral blameworthiness, respectively. More specifically, the objection is that the reasons for which an agent acts are generally irrelevant to moral wrongness, but not to moral blameworthiness. I will call this the *objection from the differential relevance of motivating reasons*. Since motivating reasons for action are closely related to the intentions with which agents act, the objection can equally be formulated in terms of the (ir)relevance of

intent to moral wrongness and moral blameworthiness.[10] While this objection might in principle be applied to both Corollary 1 and Blameworthy only if Wrong, it poses more serious problems for Blameworthy only if Wrong. I will explain why the objection does not seriously threaten Corollary 1, before considering its application to Blameworthy only if Wrong.

Corollary 1, recall, is the claim that if it is morally wrong for an agent to φ, then φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them. If the objection from the differential relevance of motivating reasons is to threaten this claim, there must be cases in which agents perform unexcused morally wrong actions but are not morally blameworthy because they are motivated by good reasons.[11] Of course, since the action in question is morally wrong, there is a limit to how good these reasons can be. They cannot provide moral justifications for these actions. But this makes it hard to see how these agents could fail to be morally blameworthy for performing them without a moral excuse. An example that may help to illustrate this is the decision of some wealthy parents to send their children to expensive private schools. This decision is typically motivated by the understandable and laudable desire to want what is best for their children. But some think that if there are good enough state schools in the area, it is morally wrong for wealthy parents to send their children to private schools on grounds of unfairness (Swift 2004). Suppose this is correct, and suppose further that this desire is not so strong or uncontrollable as to give these parents a moral excuse. Should we then conclude that although it is morally wrong for wealthy parents to send their children to private schools, they are not morally blameworthy for doing so because they are motivated by the desire to want what is best for their children? Surely not: if this reason is not good enough to morally justify

[10] To be more precise, I take the intentions with which an agent performs an action to depend on the reasons for which they perform it.

[11] Let me clarify what I mean by 'being motivated by good reasons'. I am envisaging cases in which agents know that they are acting morally wrongly but nonetheless act for genuine (and fairly weighty) normative reasons.

their action, being motivated by it is not enough to let the wealthy parents off the hook for feeling guilty, or make it inapt for less well-off parents and children (and third-parties) to feel resentment (and indignation).

The objection from the differential relevance of motivating reasons has much more force against Blameworthy only if Wrong (an agent is morally blameworthy for φ-ing only if φ-ing is morally wrong). Here are two examples that illustrate this:

> *Bad Trolley*: A runaway trolley is headed towards five people. If Debbie does not turn the trolley, it will hit and severely injure them. If she does turn it, it will veer off and severely injure one person, Eric. Debbie turns the trolley, but not to save the five from harm. Rather, her motivating reason for turning the trolley is that it will harm Eric, who is her enemy. (cf. Kamm 2001: 156)

> *Bad Defender*: Tom, a villainous aggressor, will kill Laura's daughter if Laura does not severely injure him first. Laura severely injures Tom, but not in order to save her daughter, to whom she is indifferent. Rather, her motivating reason for severely injuring Tom is that she can thereby gain revenge for some perceived wrong he did to her.[12]

In each of these cases, it might be argued that the agents act morally permissibly despite being motivated by bad reasons. But their bad motivations are surely somehow morally significant, and it might be thought that this significance consists in the agents being morally blameworthy for performing these actions. To endorse both of these claims, however, would be to give up on Blameworthy only if Wrong: an agent is morally blameworthy for φ-ing only if φ-ing is morally wrong.

[12] Capes (2012: 428-431) discusses a similar case. In his version of the case, Laura does not know that Tom will kill her daughter if she does not kill Tom first. I discussed objections to Corollary 2-S arising from the differential relevance of beliefs and evidence to moral wrongness and moral blameworthiness in Section 3. To keep things focussed, I have removed this feature of the case to concentrate exclusively on Laura's motivating reasons.

Whether motivating reasons are relevant to moral wrongness is a highly controversial issue in normative ethics. In particular, since the intentions with which an agent acts depend on their motivating reasons, denying that motivating reasons are generally relevant to moral wrongness entails rejecting the Doctrine of Double Effect.[13] Given the neutrality desideratum defended in 4.1, a defender of MB has strong reasons to remain neutral on this issue if they can. They can do this by arguing for a pair of conditional claims: *if* motivating reasons are relevant to the moral wrongness of an agent's φ-ing, *then* they are relevant (in the same way) to the agent's moral blameworthiness for φ-ing; and *if* motivating reasons are generally irrelevant to the moral wrongness of an agent's φ-ing, *then* they are generally irrelevant to the agent's moral blameworthiness for φ-ing. In other words, moral blameworthiness is sensitive to whether motivating reasons are morally relevant or irrelevant. In the remainder of this section, I will defend these conditional claims.

Perhaps the best-known defender of the claim that agents can be blameworthy for performing permissible actions when they are motivated by bad reasons is T. M. Scanlon (2008).[14] Scanlon's reasons for accepting this claim stem from his understanding of blame and blameworthiness. I outlined Scanlon's views on blame in the last section. In his view, for an agent to be blame*worthy* for φ-ing is for φ-ing to show 'something about the agent's attitudes toward others that impairs the relations that others can have with him or her' (128). If, as Scanlon argues, motivating reasons are generally irrelevant to moral wrongness, then it follows from this understanding of blameworthiness that responsible agents who perform permissible actions for bad reasons are blameworthy for these actions, since these actions reveal that they have bad attitudes toward others. As I have emphasised throughout, MB does not analyse moral

[13] The Doctrine of Double Effect is formulated in various ways. On one formulation, it holds that 'it can be permissible to bring about bad effects, including the deaths of innocent people, provided that they are not intended either as an end or as a means but are unavoidable and proportionate side effects of the pursuit of good ends' (McMahan 2009: 345).
[14] See also Thomson (1991: 295) and Capes (2012: 428-431).

wrongness in terms of just *any* kind of blame. Rather, it analyses moral wrongness in terms of what I have been calling 'moral blame' – guilt, resentment, and indignation. It is perfectly consistent with MB to grant that there are kinds of blame and blameworthiness that do not line up very closely with moral wrongness. To see whether the objection to MB from the differential relevance of motivating reasons is compelling, we will need to look closely at how we can determine the proper objects of *moral* blameworthiness. Taking a close look at this will defang this objection by showing that moral blameworthiness is sensitive to whether motivating reasons are morally relevant or irrelevant.

Guilt, resentment, and indignation, like emotions generally, are intentional: they are directed toward or about objects. This leads to a natural way of understanding the proper objects of moral blameworthiness. For an agent to be morally blameworthy for φ-ing, on this approach, is for the agent to be a fitting target of guilt, resentment, and indignation *about* their φ-ing. As we saw in Chapter 2, all of these emotions have important ties to making amends. Guilt motivates agents to make amends through such things as apologies and offers of compensation, and resentment and indignation motivate agents to make offenders hold themselves accountable by feeling guilty (for the right reasons) and making amends. Given the close ties between moral blame and making amends, one way of approaching the question of what it is fitting to feel guilt, resentment, and indignation about is by asking what it is appropriate to make amends for through apologies and offers of compensation.[15]

I will focus especially on apologies. An important feature of apologies is that a good apology not only acknowledges responsibility for wrongdoing but also shows that the wrongdoer understands why their actions were unacceptable (N. Smith 2005: 479-480). At

[15] While this is a helpful approach, it should be emphasised that it is only a rough heuristic. Agents can have reasons for apologising to others and offering compensation for actions that they are not morally blameworthy for performing. For instance, compensation might be due because it would be unfair for the harmed agent to bear the costs of the harmful action on their own.

least if the wrong in question is relatively serious, we would think there was something amiss if a wrongdoer simply said 'I'm sorry' or 'I apologise' without any further elaboration (Murphy 2012: 167). More is required, and one of these further requirements is that the wrongdoer demonstrates understanding of the moral reasons there were not to perform the action in question. This understanding may emerge in the course of a longer interaction of which apology is one element. Blaming interactions often follow a characteristic pattern, in which an accusation of fault is followed by a justification, excuse, or apology from the accused; this may lead in turn to withdrawal of the accusation or forgiveness on the part of the accuser. In the course of this interaction, one issue that is often disputed and in ideal cases settled is the nature of the morally wrong action the accused performed – why it was objectionable, and how seriously wrong it was.

The fact that good apologies involve demonstrating moral understanding suggests that moral blameworthiness will prove sensitive to whether motivating reasons are morally relevant or irrelevant. To see this, let us return to *Bad Trolley* and *Bad Defender*. If the bad reasons for which Debbie and Laura act make their actions morally wrong, then they have something to apologise *for*: the content of their apologies for φ-ing, insofar as they are good apologies, will include reference to the moral significance of their motivating reasons. In contrast, if motivating reasons are generally irrelevant to the moral wrongness of an agent's φ-ing, then it is difficult to see how Debbie and Laura could be morally blameworthy *for φ-ing* (though, as we will see shortly, this is consistent with thinking that they are morally blameworthy for their bad motivations). This is because their apologies would necessarily be bad ones: they could not demonstrate moral understanding, because the actions they performed are *ex hypothesi* not morally wrong and the motivations from which they acted are *ex hypothesi* not moral wrong-makers. In summary, once we distinguish moral blameworthiness from other kinds of blameworthiness, the force of the objection from the differential relevance of motivating

reasons vanishes. Moral blameworthiness is sensitive to the moral relevance or irrelevance of motivating reasons.

Before concluding this section, I will consider two (related) objections to the argument I have defended. Here is the first. My approach has been to distinguish among kinds of blame and blameworthiness and to argue that what I am calling *moral* blame and *moral* blameworthiness are sensitive to whether motivating reasons are morally relevant or irrelevant. It might be granted that there are other kinds of blame besides moral blame such that an agent can be blameworthy for φ-ing in these ways if φ-ing is permissible but motivated by bad reasons. Being blameworthy in Scanlon's sense might be one such kind, and perhaps there are others (for instance, kinds of blame and blameworthiness that primarily concern the character of the blamed agent). But, it might be urged, agents are also *morally* blameworthy if they perform permissible acts for bad reasons. In *Bad Trolley* and *Bad Defender*, Debbie and Laura are motivated by the harm of Eric and Tom, and it might be argued that this must warrant *some* resentment on the part of Eric and Tom and *some* guilt on the part of Debbie and Laura.

Someone who defends this claim needs to show, first of all, that various other kinds of blame do not exhaust the responses that are apt in these cases. But even if this claim is established, the objection still does not threaten MB. This is because a defender of MB can allow that it is apt for agents who perform permissible actions for bad reasons to feel guilt and for others to feel resentment and indignation so long as the proper object of these emotions is not the actions they perform, but the motivations from which they perform them. This would entail by Blameworthy only if Wrong that these motivations are morally wrong (provided that the agents are not exempt from moral responsibility), but this is consistent with the claim that the actions they prompt are morally permissible. Someone who holds this combination of claims thinks that we can perform the right actions for the wrong reasons. Clearly, this is a consistent (even common-sense) position. So, even if it is argued that agents who act morally

permissibly for bad reasons are morally blameworthy for *something*, it does not follow that they are morally blameworthy for their *actions*.

Here is a second (related) objection. Again, it targets my case for the claim that *if* motivating reasons are generally irrelevant to the moral wrongness of an agent's φ-ing, *then* they are generally irrelevant to the agent's moral blameworthiness for φ-ing. The objection alleges that I have ignored the most compelling reason for thinking that an agent's moral blameworthiness for φ-ing depends on their motivating reasons for φ-ing. This is the Strawsonian idea that moral blameworthiness is responsive to the *quality of will* of the morally blameworthy agent – to whether their action displays good will, ill will, or indifference toward others (Strawson 1962/2003).[16] It might be argued that, in acting for bad reasons, agents display poor quality of will and are hence morally blameworthy for performing these actions, even if these actions are morally permissible.

To evaluate this objection, we need a clear understanding of the idea that moral blameworthiness is a matter of quality of will. I will work with David Shoemaker's recent and well worked-out account of the quality of will implicated by 'accountability blameworthiness' (he takes guilt, resentment, and indignation to be a subset of the reactions licensed by this kind of blameworthiness) (Shoemaker 2015). In Shoemaker's view, accountability blameworthiness depends on an agent's quality of regard, and this is determined by three things: (a) whether the agent perceives relevant facts about another's normative perspective as putative reasons; (b) whether the agent judges that these facts are actually reasons; and (c) whether the agent gives these reasons appropriate weight in their deliberations (98).[17]

[16] Capes (2012) and Kauppinen (2017) argue on this ground that agents can be morally blameworthy for performing permissible actions for bad reasons.

[17] Actually, Shoemaker argues that this is only one kind of quality of regard ('evaluational regard'); the other he calls 'emotional regard' (99-103). Emotional regard is a matter of whether one's emotional responses are in tune with the fortunes or misfortunes of others, especially others with whom one stands in close personal relationships. Given that the quality of regard relevant to the question of the bearing of motivating reasons on moral blameworthiness is evaluational regard, I ignore emotional regard in the main body of the text.

Given this understanding of quality of will, it seems clear that agents who act for bad reasons display a poor quality of will even if these actions are morally permissible. In being motivated by bad reasons, these agents do not give them appropriate weight in their deliberations – they give them weight as reasons *for* the actions they perform, when they are in fact reasons *against*. But this does not mean that we should concede the claim that the agents are morally blameworthy for these actions. As I stressed earlier, given the close ties between moral wrongness and making amends, we should be very reluctant to claim that agents can be morally blameworthy for performing permissible actions that are motivated by bad reasons. And we saw that it is possible to accommodate the claim that these agents are morally blameworthy for *something* by claiming that their motivations are morally wrong. Taking this line, we could agree with Strawsonians that moral blameworthiness depends on quality of will while claiming that the objects of moral blameworthiness include not only actions but also motivations.

4.6 The Consequentialist Tradition

In this section, I consider challenges to the extensional adequacy of MB arising from the consequentialist tradition. Not every version of consequentialism creates trouble for MB. For example, *rule*-consequentialist moral theories can be combined with MB without difficulty, and defenders of these theories, such as R.B. Brandt and Brad Hooker, have tended to be sympathetic towards MB (Brandt 1979: 163-176; Hooker 2000: 72-75). I will focus on two versions of consequentialism that might seem to sit less comfortably with MB: maximising, agent-neutral versions of act-consequentialism (henceforth, simply 'act-consequentialism') and maximising, agent-neutral versions of global consequentialism (henceforth, simply 'global

consequentialism'). I will argue that, insofar as these views generate extensional challenges to MB, MB can meet them.

4.6.1 Act-Consequentialism

MB does not, by itself, entail the falsity of act-consequentialism. Combining act-consequentialism with MB yields the following claim:

> *Blameworthy iff Non-Optimific*: Φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, if and only if φ-ing fails to maximise value.

Blameworthy iff Non-Optimific is a coherent view. Since act-consequentialism applies to all actions, it applies to the social actions of overtly blaming others and imposing sanctions on them. Hence, act-consequentialism implies that we are morally required to blame others overtly and impose sanctions on them for performing actions that were not morally wrong, if the consequences of doing so are better than the consequences of any alternative option open to us. But this is not inconsistent with MB. As we saw in Chapter 1, MB understands moral blame in terms of the emotions of guilt, resentment, and indignation. MB does not by itself have any implications for the question of when overt blame is morally wrong, permissible, or required. There may be a tension in claiming that the moral blaming emotions can be fitting towards actions that we are morally required not to blame overtly, and *vice versa*, but this tension does not amount to an inconsistency or contradiction. So, MB is compatible with act-consequentialism.

However, this is not the end of the discussion. The trouble is that Blameworthy iff Non-Optimific, although coherent, is very implausible, as the following case shows:

> *Altruistic Alfie*: Alfie donates a large amount of his salary to effective aid agencies, and he volunteers for local charities at the weekend. While Alfie's actions do a lot of good, he would do more good if he donated *almost all* of his salary to effective aid agencies, and spent *almost all* of his spare time volunteering for local charities. Alfie would be a lot less happy if he lived this way, but the overall good he would produce would be greater than the overall good he produces currently.

Intuitively, Alfie does not violate standards such that, if he violated those standards without a moral excuse, he would be morally blameworthy for violating them (cf. McElwee 2017). Indignation towards Alfie for not doing more seems entirely out of place; on the contrary, his current lifestyle seems admirable. But since Alfie's actions are non-optimific, this stands in tension with Blameworthy iff Non-Optimific. Now, act-consequentialism and MB jointly entail Blameworthy iff Non-Optimific. So, if we have very good reasons for rejecting Blameworthy iff Non-Optimific, then it seems we should reject either act-consequentialism or MB. But this creates trouble for MB: it seems that act-consequentialists might well argue, on the strength of the arguments in favour of their view, that we should reject MB.

Against this, I will argue that we should resolve this tension by rejecting act-consequentialism *as a theory of moral wrongness* – but this is compatible with going on to re-interpret and affirm act-consequentialism as a claim about a different subject matter, such as *what we have most moral reason to do*. Methodologically, it seems best, *ceteris paribus*, to resolve the tension between MB and act-consequentialism in a way that accommodates as well as possible all of the features that make these views plausible (cf. Dorsey 2013: 29). If we reject Blameworthy iff Non-Optimific, we cannot accept both MB and act-consequentialism as they stand. But it might be possible to respect the features that make these views plausible while re-interpreting one or both of these views in a way that eliminates the conflict between them. It seems there are various ways in which we could re-interpret act-consequentialism that

eliminate the conflict between it, MB, and the rejection of Blameworthy iff Non-Optimific. For example, we might understand it as making the following claim:

Most Moral Reason: We always have most moral reason to maximise value.[18]

This claim is compatible with MB and the rejection of Blameworthy iff Non-Optimific. This is because it is conceptually possible that failing to act as we have most moral reason to act is not always morally wrong (Darwall 2017: 6-7; Chappell 2020: 504-508). For example, consider the moral theory that Samuel Scheffler defends in *The Rejection of Consequentialism* (1982). Scheffler accepts Most Moral Reason, but argues that we have an agent-centred moral permission to do less than best when producing the most good would impose sufficiently great personal costs on us. Hence, in Scheffler's view, it is possible for someone to have most moral reason to maximise value even though failing to do so would not be morally wrong, because of this agent-centred moral permission.

Insofar as Most Moral Reason respects many of the features that make act-consequentialism plausible in the first place, we might resolve the conflict between act-consequentialism, MB, and the rejection of Blameworthy iff Non-Optimific by accepting Most Moral Reason instead of act-consequentialism. Moreover, there is a strong case for thinking that Most Moral Reason *does* respect these features. For example, Scheffler argues that an important part of the appeal of act-consequentialism is that it embodies a *prima facie* attractive conception of rationality that he calls 'maximising rationality':

> The core of this conception of rationality is the idea that if one accepts the desirability of a certain goal being achieved, and if one has a choice between two options, one of

[18] Norcross (2006), Chappell (2020), and McElwee (2021) all argue that Most Moral Reason is the most plausible form of (maximising) act-consequentialism.

> which is certain to accomplish the goal better than the other, then it is, *ceteris paribus*, rational to choose the former over the latter. (1988: 414; cf. Pettit 1991)

Most Moral Reason can be seen as embodying this conception of rationality, at least if we understand rationality in terms of responsiveness to reasons (and Scheffler himself seems to have some such 'substantive' notion of rationality in mind). This suggests that moving from act-consequentialism to Most Moral Reason need not lead to abandonment of the features that make act-consequentialism plausible.[19]

In contrast, it seems harder to reject Blameworthy iff Non-Optimific and hold on to act-consequentialism while respecting the motivations for accepting MB. These include the pre-theoretical attractiveness of the idea that there is a close link between moral wrongness and moral blameworthiness, the ability of MB to explain the connection between being a morally responsible agent and being subject to moral requirements, and the possibility of generalising MB to provide a unified account of moral and non-moral requirements. So, giving up MB would incur greater costs than giving up act-consequentialism. If we give up act-consequentialism, we could move to a similar view that accommodates much of what makes act-consequentialism plausible, but this seems harder to do in the case of MB. For this reason, the conflict between act-consequentialism, MB, and the rejection of Blameworthy iff Non-Optimific is best resolved by giving up act-consequentialism. I should emphasise that I do not mean to endorse Most Moral Reason overall. My claim is only that, as far as the conflict between MB, act-consequentialism, and the rejection of Blameworthy iff Non-Optimific is concerned, we do better to reject act-consequentialism in favour of Most Moral Reason than to reject MB.

[19] See McElwee (2019) and Chappell (2020) for further discussion.

Before moving on to discuss global consequentialism, it is worth relating the preceding discussion of act-consequentialism back to the neutrality desideratum defended in Section 1. There, I argued that defenders of analyses of normative concepts should typically be very wary of taking on first-order commitments. This is because defenders of such analyses are committed not only to claiming that first-order views that directly conflict with their analyses are false, but that they can be shown to be false solely by appealing to the platitudes associated with the relevant normative concepts. Now, MB does not directly entail the falsity of act-consequentialism. However, it does directly entail the falsity of act-consequentialism combined with the rejection of Blameworthy iff Non-Optimific – and this combination of claims is itself a first-order view. This places a heavy dialectical burden on MB, since defenders of this view are committed to arguing that this combination of claims can be shown to be false solely by appealing to the platitudes associated with MORAL WRONGNESS. However, on reflection, it seems that this dialectical burden can be discharged. We saw in Chapter 1 that there is a wide range of seeming platitudes surrounding MORAL WRONGNESS that MB is well-placed to capture. Moreover, even if we go on to reject Blameworthy iff Non-Optimific and hence give up act-consequentialism, there is a closely related view – Most Moral Reason – that accommodates much of what makes act-consequentialism plausible.

4.6.2 Global Consequentialism

Let me turn next to global consequentialism. An influential statement of the view is due to Michael Smith and Philip Pettit:

> Global consequentialism identifies the right x, for any x in the category of evaluands – be the evaluands acts, motives, rules, or whatever – as the best x, where the best x, in turn, is that which maximises value. (2000: 121)

Interpret 'right x' as 'morally required x'. Then global consequentialism, as applied to emotions, claims:

> *Global Consequentialism (Emotions)*: An agent is morally required to feel an emotion, E, if and only if E maximises value.

We saw in Chapter 1 that MB can be formulated in terms of moral requirement:

> *MB-Requirement*: An agent is morally required to φ if and only if (Def) not φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them.

Global Consequentialism (Emotions) and MB-Requirement are a coherent combination of views, but they yield some odd results. One such result is that we may be morally required to feel an emotion that it would not be fitting to feel, and that we may be morally required not to feel an emotion that it would be fitting to feel. Even more oddly, this combination of views gives the result that it may sometimes be fitting to feel a moral blaming emotion, E-2, *about* fittingly feeling another moral blaming emotion, E-1, if feeling E-1 did not maximise value. For example, suppose that it would be fitting to feel guilty about φ-ing, but feeling such guilt would be counter-productive. Then, according to Global Consequentialism (Emotions), the agent is morally required not to feel guilty about φ-ing. But now suppose the agent does feel guilty about φ-ing, and, moreover, that they lack a moral excuse for feeling this emotion. Then, according to MB-Requirement, it is fitting for them to feel guilty about fittingly feeling guilty about φ-ing. This is an odd result. But it is not in itself an incoherent claim.[20] MB-Requirement,

[20] Indeed, claims of the general form that it is fitting to feel an emotion, E-2, about fittingly feeling another emotion, E-1, are not only coherent, they are sometimes very plausible. For example, suppose that someone fittingly feels happy about X after coming out of a long period of depression. It seems plausible to claim that it would be fitting for them to feel happy about fittingly feeling happy about X. What makes the example concerning guilt odd is that guilt is an emotion of self-criticism.

then, does not by itself entail that global consequentialism is false or incoherent. To this extent, it is neutral with respect to it.

It might be wondered, however, whether these results of Global Consequentialism (Emotions) and MB-Requirement are so implausible that we end up in a dialectical situation that is similar to the one we encountered with act-consequentialism and MB: *viz.*, that we have very good reasons to reject one of these views, even though they are not directly contradictory. However, the dialectical situations are importantly different. The key point is that global consequentialists are already committed to a wide range of claims that are odd in a very similar way, in giving normative verdicts that are not directly contradictory but nevertheless in apparent tension. For example, they are committed to claiming that we may be morally required to act from a morally wrong set of motives, and *vice versa*. If global consequentialists are able to show that these claims, despite their oddness, do not give us sufficient reasons for rejecting their theory, then we should expect their arguments for this conclusion to generalise to the odd claims introduced in the previous paragraph. For example, consider the following passage from Julia Driver:

> …judging normative ambivalence to be *appropriate* rather than problematic goes a long way toward mitigating the air of paradox. One should do what maximises the good but also be the sort of person who fails to maximise the good when that failure is the result of dispositions it is important for people to have – dispositions that are crucial to human happiness, for example… …It is not at all surprising that guidance might be mixed, because guidance about what to *do* is different from guidance about what to *be*. (2014: 175)

If Global Consequentialism captures the normative ambivalence appropriate in cases where the best set of motives would issue in an action that does not have the best consequences, then we

might similarly argue that Global Consequentialism and MB-Requirement capture the normative ambivalence appropriate in cases where experiencing a fitting moral blaming emotion would have bad consequences. It seems, then, that the odd claims generated by Global Consequentialism and MB-Requirement do not affect the overall plausibility of Global Consequentialism. If the odd claims Global Consequentialism gives rise to *by itself* do not create insuperable problems for the theory, then neither will the very similar odd claims that result when it is conjoined with MB-Requirement.

I have been considering Global Consequentialism as a theory of moral requirements. Another way to develop Global Consequentialism is to apply consequentialism to other forms of normative assessment besides deontic moral assessment (Driver 2014: 174-175). For example, someone might defend consequentialist theories of virtues, vices, and – relevantly for our purposes – *fittingness*. Applied to fitting emotions, this yields:

> *Fitting Emotions Consequentialism*: It is fitting to feel an emotion, E, if and only if feeling E would maximise value.

Combining *Fitting Emotions Consequentialism* with MB gives us:

> It is morally wrong for an agent to φ if and only if φ-ing violates standards such that, if the agent violated those standards without a moral excuse, morally blaming them for violating them would maximise value.

This claim is very implausible. Consider the following case:

> *Infidelity/Evil Demon*: Zeno breaks an uncoerced promise he made to Kelly without a moral excuse. Keeping the promise would have been mildly inconvenient, but this was entirely foreseeable when it was made, and breaking it has disastrous consequences. An evil demon credibly threatens to destroy the planet if anyone morally blames Zeno for breaking his promise.

Clearly, breaking the promise was morally wrong. But the combination of Fitting Emotions Consequentialism and MB mistakenly gives the verdict that this was not morally wrong, because morally blaming Zeno for breaking the promise would not maximise value. So, if defenders of MB are to avoid giving very implausible verdicts in cases such as *Infidelity/Evil Demon*, they must reject Fitting Emotions Consequentialism.

There are compelling grounds for rejecting Fitting Emotions Consequentialism. Indeed, it is conceptually false. In Chapter 1, I introduced fittingness by pointing to a range of terms that are often used to express this relation (such as 'apt' and 'merited') and noting that many terms for emotions and other intentional states have associated evaluative terms formed by suffixation, such that the state in question is fitting if and only if its object falls under its associated evaluative term (e.g., 'admiration'/'admirable', 'amusement'/'amusing', 'shame'/'shameful', 'desire'/'desirable'). With this gloss on fittingness in hand, consider the following example:

> *Evil Demon*: An evil demon credibly threatens to torture you unless you admire him. (Rabinowicz and Rønnow-Rasmussen 2004: 407)

Other things equal, Fitting Emotions Consequentialism gives the result that it is fitting to admire the demon. However, the claim that it is fitting to admire the demon because admiring him would maximise value seems conceptually false. Imagine that someone claims that the evil demon is admirable. They do not mean that the demon is admirable because he is powerful or cruel – they think that there is nothing admirable about possessing these qualities. Rather, they claim that he is admirable because, given the fact that he will torture you if you do not admire him, admiring him maximises value. This seems conceptually false. Apparently making this claim is evidence that the agent lacks the concept ADMIRABLE, and will tend to produce a feeling of bafflement in competent users of this concept. Given that I introduced FITTINGNESS

partly by reference to terms such as ‘admirable’, this suggests that Fitting Emotions Consequentialism is a conceptual falsehood.

However, while we should reject Fitting Emotions Consequentialism, there are other consequentialist accounts of emotional fittingness that are not vulnerable to this critique – and which, moreover, are compatible with MB. I will focus in particular on consequentialist accounts of the fittingness of moral blame. Instead of defending a *direct* consequentialist account of the fittingness of these emotions, like Fitting Emotions Consequentialism, it is possible to defend an *indirect*, or ‘two-level’, version of consequentialism about moral blameworthiness (e.g., Vargas 2013; Miller 2014). According to:

> *Indirect Consequentialism about Moral Blameworthiness*: An agent is morally blameworthy for φ-ing if and only if they are a fitting target of moral blame for φ-ing according to the optimific code of rules.

By ‘the optimific code of rules’, I mean the code of rules such that, if (almost) everyone internalised it, things would go best. Indirect Consequentialism about Moral Blameworthiness is not threatened by cases analogous to *Evil Demon*. For example, in *Infidelity/Evil Demon*, Fitting Emotions Consequentialism implies that Zeno is not morally blameworthy because morally blaming him for breaking the promise would not maximise value, but Indirect Consequentialism about Moral Blameworthiness has no such implication. On this view, whether Zeno is morally blameworthy depends not on the consequences of morally blaming him, but rather on whether he is a fitting target of moral blame for φ-ing according to the optimific code of rules.

MB is compatible with Indirect Consequentialism about Moral Blameworthiness (cf. Miller 2014: 19). Interestingly, combining MB with this theory yields a kind of rule-consequentialist theory of moral wrongness:

> It is morally wrong for an agent to φ if and only if they would be a fitting target of moral blame for φ-ing according to the optimific code of rules.[21] (cf. Hooker 2000: 72-75)

This is not an uncontroversial claim. But, unlike the result of combining MB with Fitting Emotions Consequentialism, it is not an obviously very implausible claim. So, while defenders of MB are committed to rejecting Fitting Emotions Consequentialism on pain of being vulnerable to clear counter-examples, they are not similarly committed to rejecting Indirect Consequentialism about Moral Blameworthiness. In sum, consequentialist accounts of fitting moral blame do not pose insuperable difficulties for MB. Direct accounts, such as Fitting Emotions Consequentialism, are conceptually false, and indirect accounts, such as Indirect Consequentialism about Moral Blameworthiness, are compatible with MB.

4.7 Conclusion

In this chapter, I discussed and rejected a range of extensional objections to MB. I argued that, with the exception of extensional challenges arising from the Objectivism/Subjectivism debate, no changes need to be made to MB to meet these objections successfully. Moreover, defenders of MB can meet these extensional challenges while remaining largely neutral on the first-order debates at issue.

[21] I assume that the set of rules for when agents are fitting targets of moral blame includes rules concerning moral excuses. Hence, the clause concerning moral excuses in MB can be omitted.

Chapter 5

Circularity Objections

This chapter defends MB against the charge that it is viciously circular. The chapter is entitled circularity *objections* to MB, rather than *the* circularity objection, since, as we will see, there are various considerations that might lead one to worry that MB is viciously circular.

The most straightforward circularity objection to MB is generated by pairing the following account of what it is for emotions to be fitting with the following claim about the nature of indignation, the central moral blaming emotion:

> *Emotional Fittingness as Accurate Representation*: For an emotion to be fitting is for it to involve an accurate representation of its object.[1]
>
> *Deontic Moral Content*: Indignation involves a judgment or other attitude including the content: *it is morally wrong for the agent to φ*.

If we plug these claims back into MB, we get the following: for it to be morally wrong for an agent to φ is for φ-ing to violate standards such that, if the agent violated those standards without a moral excuse, the pair <agent, φ-ing> would be accurately represented as instantiating the moral wrongness relation (among other relations). This is clearly circular.

As we saw in the Introduction, some neo-sentimentalists defend explicitly circular analyses, but deny that the circularity is vicious (McDowell 1985; Wiggins 1987; Tappolet 2016). In contrast, my aim is to defend MB as a non-circular analysis. To do so, I need to argue

[1] See, e.g., De Sousa (2002: 249-251); Graham (2014: 392-393); Rosen (2015b: 70-71); and Tappolet (2016: 87).

against either Emotional Fittingness as Accurate Representation or Deontic Moral Content. I argue that we have good reasons to reject both.

Section 5.1 argues against Emotional Fittingness as Accurate Representation. Emotions are complex, and typically involve various elements. As well as involving representations that can be assessed for accuracy, emotions typically motivate their subjects in characteristically urgent ways, involving direction of attention and bodily preparation. I argue that the fittingness of an emotion as a whole is a function of the fittingness of both its representational and motivational aspects (I call this 'The Hybrid Model of Emotional Fittingness'). My central argument for The Hybrid Model of Emotional Fittingness is that it is only by recognising the contributions that both of these aspects make to emotional fittingness that we can make sense of the proportionality constraint on emotional fittingness. By the 'proportionality constraint' on emotional fittingness, I mean, roughly, the condition that, to be fitting, an emotion cannot be too strong or too weak.

Section 5.2 argues against Deontic Moral Content. I distinguish between two ways of denying Deontic Moral Content. The first way is to deny that indignation has any normative content whatsoever. The second way is to deny that indignation has any moral content, but leave open the possibility that it may contain some (non-moral) normative content. I pursue the second approach, and defend the viability of this kind of approach to defending neo-sentimentalist analyses against an important general challenge raised by Justin D'Arms and Daniel Jacobson (e.g., 2003: 127-128, 2023: 91). The challenge is that, if emotions turn out to have independently intelligible normative content, fitting emotions drop out of the resulting analyses as redundant, since we could just analyse the concepts and/or properties under consideration directly in terms of the normative content of the emotional responses. I argue that this challenge depends for its force on Emotional Fittingness as Accurate Representation.

If we give up this account of emotional fittingness in favour of The Hybrid Model of Emotional Fittingness, the challenge loses its force.

Section 5.2 also defends an account of what is involved in making amends. It might be worried that pairing MB with The Hybrid Model of Emotional Fittingness leaves MB open to a different circularity objection, based on the claim that the motivational goal of indignation, which is, roughly, to get offenders to hold themselves accountable for φ-ing by feeling guilty (for the right reasons) and making amends, cannot be understood independently of moral wrongness. The concern is that an adequate account of what is involved in making amends will make ineliminable reference to moral wrongness. Against this objection, I defend an account of what is involved in making amends that makes no reference to moral wrongness. This account characterises making amends in terms of the harms making amends seeks to repair, and the means by which it repairs these harms. Drawing on work by Jeffrie Murphy (1982), Pamela Hieronymi (2001), Linda Radzik (2009), and others, I focus in particular on how making amends withdraws threatening messages that are typically sent out by unexcused violations of certain standards.

5.1 Emotional Fittingness

I begin (in 5.1.1) by introducing and criticising Emotional Fittingness as Accurate Representation ('EFAR' for short). I argue that EFAR fails to give an adequate account of the proportionality condition on emotional fittingness. This failure stems from EFAR trying to account for emotional fittingness exclusively in terms of one aspect of emotional episodes, their representational aspect. This diagnosis of why EFAR is unsuccessful motivates a complex, hybrid account of emotional fittingness that acknowledges the contributions that different

elements of emotional episodes make to emotional fittingness. I develop an account on these lines (in 5.1.2) and defend it against various objections.

Before I start, let me make a clarification. Statements ascribing emotions such as 'Beth feels guilty about hurting Sue' typically admit of two different readings (cf. Deonna and Teroni 2012: 8). On the first reading, this statement says that Beth is right now in the thrall of guilt. So understood, the statement ascribes an emotional episode to Beth. But we might also read the statement as saying that Beth is disposed to feel episodes of guilt about hurting Sue, for example when she thinks about Sue. Understood this way, the statement ascribes an emotional disposition to Beth. In providing an account of emotional fittingness, my aim is to provide an account of the fittingness of emotional episodes. I leave open whether this account might be extended to give an account of the fittingness of emotional dispositions. In what follows, by 'emotion' I usually mean 'emotional episode' unless otherwise indicated.

5.1.1 Against Emotional Fittingness as Accurate Representation

It is widely thought that, in typical cases, emotional episodes involve various components (Goldie 2000: 4-5; Prinz 2004: 3-4; Tappolet 2016: 8; Robinson 2017; Scarantino and De Sousa 2021: §2). First, emotions involve some kind of representation of their objects. On virtually all views, emotions involve non-normative representations of their objects. For example, my fear of a rapidly approaching bear involves a representation of the bear, as well as representations of some of its non-normative features, such as its size and movement. Many philosophers argue further that emotions involve normative representations of their objects (e.g., Goldie 2000; Nussbaum 2001; Roberts 2003; Tappolet 2016). In the case of fear, a popular candidate is dangerousness: fear, it is often claimed, involves as a component a representation of its object as dangerous. Second, emotional episodes typically involve various bodily changes, such as

changes in heartrate and facial expressions. Third, emotions typically involve characteristic feelings, such as 'heated' feelings in anger. Fourth, emotions typically involve changes in attention, and, more generally, information processing. Fifth and finally, emotions typically involve a motivational component that is directed towards a characteristic goal. For example, fear typically motivates its subjects to become safe from its objects.

Now, we will see later that the claim that emotions involve different components can be misleading, insofar as it might lead us to overlook the important ways in which these components are integrated with one another. Moreover, it may be that some of the components mentioned above, properly understood, are fully explicable in terms of the others (for example, perhaps the phenomenal character of emotions can be fully explained in terms of the other elements). Nonetheless, the claim that emotions are complex phenomena will be important in what follows, and I will return to it at various points in the argument.

EFAR attempts to understand emotional fittingness in terms of only the first, representational component of emotions. It is compatible with a wide range of views about the nature of the representations involved in emotional episodes. For example, a defender of EFAR might hold that emotions involve normative judgments (Nussbaum 2001), or perceptual experiences of normative properties (Tappolet 2016), or *sui generis* normative representations (Goldie 2000; Mitchell 2020).[2] Moreover, EFAR is not essentially tied to the view that the representations involved in emotions are necessarily normative (as opposed to non-normative) (cf. Deonna and Teroni 2021: 1112). EFAR's flexibility in these respects is one source of its appeal. A further strength of EFAR is that by reducing fittingness to accurate representation, it

[2] I intend 'involved' to be understood sufficiently broadly that even philosophers who hold that emotions depend on prior representations of their objects (e.g., Deonna and Teroni forthcoming) might allow that these emotions 'involve' such representations.

reduces something that is often found to be somewhat obscure (fittingness) to a comparatively clearer and better understood category (representational accuracy).

EFAR ultimately founders on the issue of proportionality.[3] To be fitting, the strength of a response must be proportional to the value of its object (cf. D'Arms and Jacobson 2000b: 73-74). For example, suppose that Josephine feels strongly ashamed of something that is only a little shameful. While Josephine's shame gets something right – it is directed towards something that really is shameful – it fails to be fitting, because the strength of her shame is disproportional to the degree of shamefulness of its object. To give a different example, suppose that Mike feels only weakly guilty about brutally assaulting someone. Mike's guilt is not fitting, because brutal assaults merit much stronger guilt than this.

It bears emphasising that 'strength', as used here, is a term of art. It is pre-theoretically plausible that fitting emotions are subject to some kind of proportionality condition. Moreover, as we will see shortly, we have some intuitions about the components of emotions that are relevant to whether this condition is met. But we should not take for granted that everything we might reasonably call an aspect of the 'strength' of an emotion is relevant to its fittingness. For example, the felt intensity of an emotion is partly a function of the feelings of bodily changes it involves, but it is an open question how, if at all, such feelings contribute to emotional fittingness. So, 'strength' should be understood as meaning 'aspects of an emotion that bear on whether it meets the proportionality condition on emotional fit'.

A defender of EFAR, it seems, must attempt to account for proportionality by appealing to representational content: an emotion that represents a certain normative property as being

[3] A further argument against EFAR, first made by Sigrún Svavarsdóttir, is that it is undermined by the intuition that unfitting emotions involve a different kind of mistake than merely inaccurate representation (Svavarsdóttir 2014: 89-90, 101; Howard 2018: 6, 11; Naar 2021: 13609-13610; D'Arms 2022: 122). The hybrid account of emotional fittingness I defend later identifies a further mistake involved in unfitting emotions: *viz.*, an unmerited motivational state.

instantiated to a greater degree is stronger than one that represents it as being instantiated to a lesser degree.[4] For example, the strength of an episode of fear is a function of how dangerous it represents its object as being. The problem with this approach to understanding the proportionality condition on emotional fittingness is that other aspects of emotions besides their representational component are intuitively relevant to whether emotions are proportional. For example, imagine that Tony's fear accurately represents its object as being mildly dangerous, but also involves intense, exclusive attention on its object and very strong motivation to become safe from it. Even though this fear episode involves an accurate representation of its object, Tony's emotional response as a whole seems disproportional and hence unfitting. To give a different example, suppose that Beatrice's shame accurately represents its object as a very serious personal inadequacy, but involves only weak motivation towards self-improvement or concealment. Again, although this episode of shame accurately represents its object, Beatrice's emotional response as a whole seems disproportional and hence unfitting.[5]

In response, a defender of EFAR might argue that strength of attentional focus and motivation are not relevant to the 'strength' of an emotional episode in the technical sense at issue: that is, they are not aspects of an emotion that bear on whether it meets the proportionality condition on emotional fittingness. However, this response is undermined by reflection on the kinds of criticisms to which agents can be open in virtue of experiencing disproportional emotions. When we criticise someone's emotion by saying that its object is not *that* embarrassing, amusing, admirable, and so on, it seems plausible that in typical cases a large part of what is targeted by this criticism is strength of attentional focus and motivation. For example, imagine that someone commits a minor *faux pas*, but is absolutely mortified: 'I can't

[4] Perhaps a defender of EFAR might also appeal to the duration of emotional episodes.
[5] Clarke and Rawling (forthcoming) argue similarly that accounts of the fittingness of blaming emotions in terms of representational accuracy have problems accounting for proportionality. Their arguments are more complex than mine, concerning changes in the fittingness of blaming emotions over time.

stop thinking about it', she says, 'I can never show my face there again'. If we reply to her, 'It really wasn't *that* embarrassing', it seems very plausible that part of what we are targeting in characterising her degree of embarrassment as an unfitting overreaction is its attentional and motivational aspects.

In a recent paper, Justin D'Arms (2022) proposes a more promising way of alleviating concerns around proportionality for EFAR.[6] He suggests that defenders of EFAR might address these concerns by adopting an account of how emotions normatively represent their objects that he has developed in joint work with Daniel Jacobson (e.g., D'Arms and Jacobson 2022, 2023: 135-152). On this account, emotions do not normatively appraise their objects in the sense of involving, as a component, a normative belief, perceptual experience, or the like. Rather, emotions themselves, understood as syndromes including feelings, patterns of attention, and motivations, as a whole normatively appraise their objects (2022: 18, 2023: 135-152). For example, on their view, fear represents its objects as dangerous in the sense that fear as a whole, understood as a syndrome including among other things feelings of dread and motivation to avoid threats, appraises its objects as dangerous (2017: 267).

D'Arms argues that defenders of EFAR might draw on this account of emotional appraisal to capture the relevance of such things as attentional focus and strength of motivation to the proportionality of emotional responses (2022: 121). The central idea is that the degree to which emotions represent certain normative properties as instantiated is partly a function of the attentional focus and strength of motivation they involve (Ibid.). For example, on this approach, how dangerous an episode of fear represents its object as being depends in part on how tightly it focusses its subject's attention on its object and how strongly it motivates them to become safe from it. In this way, defenders of EFAR might try to hold on to the claim that the 'strength'

[6] Although D'Arms argues that defenders of EFAR may be able to handle concerns around proportionality, he ultimately rejects EFAR on the basis of the objection mentioned in fn. 3.

of an emotion, in the sense relevant to its proportionality, is a matter of the degree to which it represents certain normative properties (such as dangerousness) as instantiated.

However, there is a good reason for rejecting D'Arms and Jacobson's account of emotional appraisal. It seems common for strength of attentional focus and motivation involved in emotional episodes to fade as they become more distant in time from their particular objects (cf. Clarke and Rawling forthcoming: Sec. 6). For example, compare the resentment that someone might feel for a minor wrong done to them shortly after the offence took place, with the resentment they might feel 30 years later (Ibid.). It seems probable that the strength of attentional focus and motivation involved in the later episode of resentment will be milder than that involved in the earlier episode. But surely this need not correspond to a difference in how serious these episodes of resentment represent the wrong as being. The subject's resentment might well involve a clear-eyed view of the seriousness of this wrong in both cases; what has changed in the intervening years is that the wrong no longer exercises them as much. So, we should reject the claim that the degree to which emotions represent normative properties as instantiated is a matter of such things as the strength of attentional focus and motivation they involve, and with it the holistic account of emotional appraisal that D'Arms and Jacobson defend. To make sense of how the representational content of an emotion can vary somewhat independently of its attentional and motivational aspects, it seems that we must understand emotional appraisal as a component of emotions in its own right.

I conclude, then, that we should reject EFAR, on the basis that it fails to give a satisfactory account of the proportionality condition on emotional fit. This does not mean that appealing to representational accuracy should not play *any* role in an account of emotional fittingness. The central mistake of EFAR seems to be that it focuses exclusively on representational accuracy and ignores the contribution of other elements of emotional episodes to emotional fittingness, such as their attentional and motivational aspects. Indeed, an account

of emotional fittingness that focussed only on the latter elements would seem to go wrong in an exactly parallel way to EFAR. It seems easy to imagine cases in which the strength of attentional focus and motivation involved in an emotional episode are proportional, but the representational aspect of the emotional episode is disproportional. For example, imagine that Peter feels ashamed of his cowardice, and the attentional and motivational aspects of his shame are proportional to the shamefulness of his cowardice. But Peter's shame also involves a misrepresentation of his cowardice: it represents it as only a minor personal shortcoming, when it is in fact a serious personal inadequacy. Peter's shame as a whole seems disproportional and hence unfitting. The lesson seems to be that an adequate account of emotional fittingness needs to acknowledge that emotional episodes are complex, and needs to explain how different components of an emotional episode contribute to its fittingness.

5.1.2 The Hybrid Model of Emotional Fittingness

In the remainder of this section, I develop a hybrid account of emotional fittingness that attempts to do just this. Schematically, hybrid accounts of emotional fittingness claim that for an emotional episode to be fitting is for elements X, Y, Z, … of that emotional episode to be fitting. We can flesh out this schema by saying more about the different elements of emotions that go into it.[7] We need, first, a better understanding of the ways in which the various components of emotional episodes are integrated with one another. As we will see, the key to understanding this is to see the motivational aspect of emotional episodes as a focal point that organises many of the other aspects of emotional episodes.

[7] To flesh it out further, we would also need to say more about the nature of the fittingness relation. As I explained in the Introduction, I wish to remain neutral between the views that fittingness is normatively primitive, or else analysable in terms of some further normative category such as reasons or values. Below, I say a little about how these different approaches to understanding fittingness in general relate to the hybrid account of emotional fittingness defended here.

Emotional motivation is distinctive in having the property of 'control precedence'.[8] A motivational state has control precedence insofar as it tends to assume precedence in the control of action, attention, and information-processing. Control precedence involves various elements, which I distinguish under the headings of 'bodily preparation', 'effects on attention and information-processing', and 'interruption'. Although these elements are distinguishable, they form a unified phenomenon insofar as they all contribute to the prioritisation of the motivational state to which they belong.

I said earlier that the claim that emotions involve different components can be misleading. Although we can analytically distinguish the various components typically involved in emotional episodes, these components are integrated in significant ways. In particular, insofar as emotional motivation involves control precedence, the motivational component serves as a kind of focal point or organising principle in terms of which many of the other elements typically involved in emotional episodes can be understood – in particular, bodily changes, along with changes in attention, and, more generally, information processing. Although I will continue to talk about the 'motivational component' of emotions, from now on I intend this to refer to the distinctive motivational states typically involved in emotions. Insofar as these motivational states are distinctive in having the property of control precedence, it is misleading to *contrast* the motivational component of emotions with bodily changes and changes in information processing. All of these things contribute to control precedence and hence to the distinctive character of emotional motivation. Let me now explain control precedence in more detail.

Bodily preparation: Motivational states with control precedence prepare the body for action in service of their goals. This can involve changes in muscular tension, including

[8] The term 'control precedence' was coined by the psychologist Nico Frijda (1986, 2007). My presentation of control precedence will largely follow Scarantino (2014).

changes in facial expressions, and changes in the autonomic nervous system, such as changes in heartrate and blood flow. In fear, for example, heartrate tends to increase and blood tends to flow to large muscles; fear also tends to involve changes in facial expressions, such as raised eyelids (Scarantino 2014: 159). In this way, fear prepares us to become safe from its objects by preparing our body for quick action (such as running away from a physical threat) and increasing our field of vision.

Effects on attention and information-processing: Motivational states with control precedence affect the attention and information-processing of their subjects in service of their goals (Frijda 2007: 41; Scarantino 2014: 168-171). In guilt, for instance, our attention is typically directed towards an action we have performed or omitted, and also on actions that we could perform to make amends for it, such as apology or compensation (Tangney and Dearing 2002: 18-20).

Interruption: Motivational states with control precedence tend to interrupt ongoing pursuits and not to be interrupted themselves (Frijda 1986: 78; 2007: 28-29; Scarantino 2014: 168-171). For instance, an episode of fear may distract its subject from her absorption in a book when she becomes aware of a large dog bounding towards her and her fear will tend not to be interrupted until she has become safe.

On the hybrid model I defend, for an emotion to be fitting is for both its representational and motivational components to be fitting:

> *The Hybrid Model of Emotional Fittingness* (HEF): For an emotion, E, to be fitting is for E's representational and motivational components to be fitting.

As I said in the Introduction, I remain neutral between the views that fittingness is normatively primitive, or else analysable in terms of some further normative category such as reasons or values. Now, it is typically thought that, whatever else we say about fittingness, a belief is

fitting if and only if it is true (Chappell 2012: 688; McHugh and Way 2016: 584). I will assume that something similar applies to the representations involved in emotions (which, on some views, just are beliefs): these representations are fitting if and only if they are accurate. So, HEF takes representational accuracy to be relevant to emotional fittingness, even though it rejects the view that for an emotion to be fitting just is for it to involve an accurate representation of its object. Instead, it takes fittingness to be a distinct relation, which, in the case of the representations involved in emotions, is instantiated just in case these representations are accurate.

Let me say more about what is involved in fittingness assessment of the motivational components of emotions. As I explained above, by 'motivational component' I mean the special motivational states typically involved in emotions. These motivational states are distinctive in having the property of control precedence. By claiming that the motivational component of emotions is relevant to emotional fittingness, HEF is in a position to acknowledge the contribution that all of the following things make to emotional fittingness, and in particular to the proportionality condition on emotional fittingness: first, strength of motivation, where this is a matter of its causal power to produce action; second, content of motivation (e.g., whether an agent in feeling guilty is motivated to make amends in an insufficient/excessive way); third, the tendency of the emotion to focus attention and prepare the body in service of its goal;[9] and fourth, the tendency of the emotion to interrupt other mental states and not be interrupted itself.[10] An emotion is fitting, according to HEF, only if all of these factors are proportional to its object.

[9] It is possible that not all of the elements of control precedence are assessable in terms of fit. It might be thought that physiological changes, such as changes in heartrate and blood flow, are not the kinds of things that can be assessed for fit. If this is right, then we should not see physiological changes as relevant to HEF.

[10] It is important that HEF focuses on *tendencies* to focus attention, prepare the body, and interrupt other mental states, rather than the extent to which emotions in fact do these things. This is because the extent to which emotions in fact do these things depends on their place in the agent's overall mental life, but the place of an emotion in the agent's overall mental life is not relevant to whether it fits its object. (Cf. Berker (2022: 44-45) on the point that

It may be that some emotion-types are associated with multiple motivational goals. For example, empirical work on admiration suggests that it is associated with being motivated to praise and honour admired others, and also to emulate them in respect of their admired qualities (Schindler et al. 2013: 100-106).[11] It might be wondered whether this poses difficulties for HEF – for example, it might be thought that defenders of HEF need to find some basis for privileging one of these motivational goals with respect to fittingness assessment. However, this is not so. Insofar as there are many ways in which people can be admirable, it should not come as a surprise that different forms of motivational engagement are fitting with respect to different kinds of admired persons or objects. I might admire a sportsperson's athleticism, but it would not be fitting for me to be motivated to emulate it. But being motivated to praise and honour the sportsperson on account of their athleticism might well be fitting. Generally, insofar as emotion-types are associated with multiple motivational goals, these different motivational goals can be seen as reflecting the complexity of the normative properties that are associated with the fittingness of these emotion-types.

HEF does a better job than EFAR in meeting the proportionality condition on emotional fittingness. To see this, let us return to the example of Tony, whose fear accurately represents its object as being mildly dangerous, but also involves intense, exclusive attention on its object and very strong motivation to become safe from it. HEF, unlike EFAR, can explain why Tony's fear is not fitting. This is because it acknowledges the contribution that the motivational component of emotions makes to emotional fittingness, and emotional motivation is distinctive

fittingness is not 'alternatives-dependent'). HEF respects this point, because the place of an emotion in the agent's overall mental life affects how the relevant tendencies are manifested, but not the tendencies themselves.

[11] Schindler et al. distinguish internalisation (internalising the values or ideals exemplified in the admired other) from imitation (realising these values or ideals in your own behaviour) (102-105). I use 'emulation' as a catch-all for both internalisation and imitation. Schindler et al. also discuss a putative further motivational goal of admiration, *viz.* towards affiliating with the admired other. They write: 'affiliation with admired others rests on their recognition as models. That is, making contact with admired others allows [us] to observe what the other is doing' (102). Insofar as admiring agents are motivated to affiliate with those whom they admire in order to learn from them as models, it seems that we might bring this motivational goal under the broader goal of emulation.

in having the property of control precedence (which encompasses the attentional changes involved in emotions). The same point applies to the other example given earlier, concerning embarrassment. When someone is mortified by a minor *faux pas*, it seems very plausible that part of what makes their embarrassment an unfitting overreaction is the intensity of attention and strength of motivation involved in it. Insofar as HEF captures the contribution of these things to emotional fittingness, it is well-placed to explain why such embarrassment is disproportionate and hence not fitting.

Let me turn now to some potential difficulties with HEF. It has been argued that some emotion-types do not, or at least do not typically, motivate (Roberts 2003: 63; Tappolet 2016: 75-76). Admiration and happiness/joy have been cited as examples of this (Ibid.).[12] Moreover, even when it comes to emotion-types that typically involve a motivational component, it may be that there are particular episodes of them that do not. For example, perhaps emotions experienced in response to fictions usually lack a motivational component, and perhaps emotions experienced in response to events in the distant past often lack such a component as well (Tappolet 2016: 64-66, 73-74; but cf. D'Arms and Jacobson 2023: 114). On the supposition that there are non-motivating emotion-types, or non-motivating emotional episodes of types that are usually motivating, how should HEF deal with them?

We need to begin by distinguishing cases in which it would be a defect in an emotion that it lacked a motivational component, and cases in which this would not be a defect. If there are emotion-types that in general lack a motivational component, it would not usually be a defect of episodes of such types to fail to be motivating. Moreover, insofar as there are various

[12] Examples of non-motivating emotion-types are controversial. We have already seen that admiration is linked with motivations to praise and honour admired others, and also to emulate them in respect of their admired qualities (Schindler et al. 2013: 100-106). Psychological work on happiness/joy appears to support the claim that it also typically has a motivational component. It has been linked with being motivated to celebrate and savour the object of happiness/joy (Smith et al. 2014: 18; Watkins et al. 2018: 524).

factors that can inhibit the motivational aspect of an emotional episode of a type that is usually motivating, some of these factors may be such that it is not a defect of an emotional episode to lack a motivational component because of the presence of these factors. For example, in the last paragraph we raised the possibility that perhaps an emotional episode may fail to be motivating because it is directed at a fictional event, or an event in the distant past. In these cases, it seems that it may in no way be a defect of the emotional episode that it fails to motivate its subject. It is not a defect of my indignation at Severus Snape for his cruel treatment of Harry Potter that it fails to motivate me to get him to feel guilty (for the right reasons) and make amends, because Snape is not a real person.

The fittingness of emotions that non-defectively lack a motivational component seems to depend solely on their representational accuracy. But perhaps there could be non-motivating emotional episodes that are defective in not including a motivational component. For example, perhaps there could be episodes of guilt that are non-motivating even though the agent committed an actual wrong and it is within their power to make amends for it. In this sort of case, although the motivational component of the emotion would be inhibited, it would not be *properly* inhibited (as it is in the case of indignation towards Snape). To be fitting, such an episode of guilt would surely need to be motivating.

Putting these points together suggests the following way of supplementing HEF should there prove to be non-motivating emotion-types, and non-motivating emotional episodes of types that are usually motivating:

> *Supplement to HEF:* If an emotion, E, non-defectively lacks a motivational component, then E's fittingness depends only on its representational accuracy. However, if it is a defect of E that it lacks a motivational component, then E is not fitting – to be fitting, E would need to involve a fitting motivation.

This supplement relies on us having an intuitive grasp of when it is a defect in an emotion that it lacks a motivational component. The examples in the previous paragraph suggest that this is a reasonable assumption.

When I first introduced the idea that emotions are complex, I mentioned five aspects of emotions: (1) representations; (2) bodily changes; (3) characteristic feelings; (4) changes in attention and, more generally, information processing; and (5) goal-directed motivation. Now, we saw that the motivational states typically involved in emotional episodes encompass (2), (4), and (5), insofar as emotional motivation has the property of control precedence. Moreover, HEF explicitly mentions (1), the representational component of emotions. But so far I have said nothing about (3), characteristic feelings. But we might well think that when an emotion is a fitting response to its object this is partly because of how it feels. For example, surely part of what makes guilt a fitting response, when it is fitting, is that guilt is painful. But then it might be thought that the most plausible hybrid account of emotional fittingness will incorporate the fittingness of the affective element of emotional episodes in addition to their representational and motivational elements.

However, it is not obvious that we need to appeal to an affective element of emotions *in addition to* their representational and motivational components. Indeed, it is not clear to what extent there even *is* an additional affective element over and above these things. HEF acknowledges the contribution that all of the following components of emotions make to emotional fittingness: (1) a representational aspect; (4) changes in attention and, more generally, information processing; and (5) goal-directed motivation. Many, perhaps all, of these elements of emotions contribute significantly to their phenomenal character.[13]

[13] Feelings of (2) bodily changes also contribute significantly to the phenomenal character of emotions. Much of the phenomenal character of fear, for example, can be explained in terms of feelings associated with the bodily changes involved in fear, such as feeling your heartrate and breathing quicken, and your hairs stand on end. (For detailed discussion of how bodily feelings contribute to the phenomenal character of emotional experience, see

Some philosophers hold that element (1) – a kind of appraisal – contributes significantly to the phenomenal character of emotions. This is common among philosophers who take emotional representation to be strongly analogous to, or even a kind of, perceptual experience (Tappolet 2016). Moreover, some philosophers hold that emotional representations involve *sui generis* feelings towards values that account for the distinctive phenomenal character of emotional experiences (Goldie 2000; Mitchell 2020). Finally, it seems that another factor that can contribute to the phenomenal character of emotions is satisfaction or non-satisfaction of their motivational goals. For example, guilt is sometimes claimed to involve painful feelings of estrangement from victims (e.g., Morris 1971), and this can surely partly be explained in terms of non-satisfaction of its goal of making amends. Once we acknowledge the contributions that (1), (4), and (5) make to emotional fittingness, it is not clear how much, if any, of their phenomenal character is left over, besides feelings of (2) bodily changes. And as I mention in fn.13, it is far from clear that bodily changes are relevant to fittingness assessments of emotions. If they are, moreover, HEF can account for this. I conclude, then, that HEF does not stand in need of revision.

5.2 Against Deontic Moral Content

According to:

> *Deontic Moral Content*: Indignation involves a judgment or other attitude including the content: *it is morally wrong for the agent to φ*.

There are two ways in which we might deny Deontic Moral Content. First, we might deny that indignation involves any normative content whatsoever. I discuss this view in 5.2.1 and some

Deonna and Teroni (2017)). But it is far from clear that bodily changes are relevant to fittingness assessments of emotions. (However, cf. Na'aman 2022). Moreover, if it turns out that such feelings are relevant to this, HEF can account for this, because bodily changes are involved in control precedence.

of the challenges it faces. Second, we might deny that indignation involves any moral content, while leaving open the possibility that indignation involves some (non-moral) normative content. This is the approach I pursue. In 5.2.2, I introduce this approach and defend it against a general challenge to this way of defending neo-sentimentalist analyses against circularity objections due to D'Arms and Jacobson (e.g., 2003: 127-128, 2023: 91). The crux of the challenge is that, if emotions turn out to have independently intelligible normative content, fitting emotions drop out of the resulting analyses as redundant: we could simply analyse the concepts or properties/relations under consideration directly in terms of the normative content of the relevant emotional responses. I argue that this challenge dissolves once we reject EFAR and accept HEF instead.

In 5.2.3, I consider a further circularity objection to MB that arises from pairing it with HEF. The objection, in brief, is that the motivational goal of indignation, which is, roughly, to get offenders to hold themselves accountable by feeling guilty (for the right reasons) and making amends, cannot be understood independently of moral wrongness. The worry is that we cannot give a satisfactory account of what is involved in making amends without making ineliminable reference to moral wrongness. Against this objection, I defend an account of what is involved in making amends that does not make any reference to moral wrongness. This account characterises making amends in terms of the harms making amends seeks to repair, and the means by which it repairs these harms. Taking a leaf from work by Jeffrie Murphy (1982), Pamela Hieronymi (2001), Linda Radzik (2009), and others, I focus especially on how making amends withdraws threatening messages that are typically sent out by unexcused violations of certain standards.

5.2.1 Emotions and Normative Content

One way to reject Deontic Moral Content is to argue that indignation contains no normative content whatsoever. The claim that emotions generally, or at least a subset of emotions, do not have any normative content whatsoever has significant precedent in the philosophy of emotion. A prominent example is the James-Lange theory of emotions, according to which emotions consist in feelings of bodily changes. Fear, on this view, consists in such things as feeling your heartrate quicken, your hairs stand on end, and some of your muscles tense. Different emotion-types are distinguished in terms of the different feelings they involve. On the original version of the theory, emotions do not have any normative content whatsoever.[14]

A defender of MB who appealed to the James-Lange theory would certainly not need to worry about circularity. But they would have other problems, because it is not clear that emotions, as understood by the theory, are the kinds of mental states that can be assessed in terms of fittingness (cf. Tappolet 2023: 72-74). In general, bodily feelings, such as an itch on your scalp, or pain from a stubbed toe, are typically not the sorts of things that can be assessed in terms of fittingness, or related normative categories such as reasons. So, it seems that the original version of the James-Lange theory cannot make sense of the claim that emotions are assessable for fit. But then the James-Lange theory of emotions, as originally presented, is not suited to featuring in defences of neo-sentimentalist analyses. It is a basic constraint on such an analysis that it appeals to a theory of emotions that allows for the possibility of fittingness assessments of emotions.

[14] See Deonna and Teroni (2012: 63-66) and Tappolet (2023: 63-74) for overviews of the James-Lange theory. Subsequent theorists have tried to develop this basic approach to make room for the claim that emotions have normative content. See especially Prinz (2004), who appeals to a teleosemantic theory of representation to argue that emotions, understood as feelings of bodily changes, represent the normative properties they were 'set up to be set off by', either through evolutionary or cultural selection.

A more promising approach for the neo-sentimentalist who wishes to claim that emotions generally, or at least a subset of emotions, do not have any normative content whatsoever is to appeal to a motivational theory of emotions. According to motivational theories of emotions, emotions essentially involve distinctive kinds of motivational states. Different emotion-types are to be distinguished primarily on the basis of the different motivations they involve.[15] Thus, fear is distinctive in aiming at safety, or threat-avoidance; contempt is distinctive in aiming at social withdrawal; disgust is distinctive in aiming at physical and sensory avoidance; and so on.

Motivational theories can be developed in different ways. One example is the motivational theory developed by D'Arms and Jacobson (e.g., 2022: 15-24, 2023: 105-135), which was briefly introduced in 5.1.1. According to D'Arms and Jacobson, emotions are complex syndromes including eliciting and attenuation conditions, feelings, characteristic thoughts, patterns of attention, and motivations. D'Arms and Jacobson count as advancing a motivational theory in virtue of the special emphasis they place on the motivational element of emotions, which they regard as crucial to type-individuating emotions.

D'Arms and Jacobson's theory is distinctive in two respects. The first is its account of normative appraisal. As we saw earlier, D'Arms and Jacobson hold that emotions *as a whole* normatively appraise their objects. Now, this might be thought to show that they do not really hold that emotions have no normative content whatsoever. However, there is a sense in which they do hold this. Their account of emotional appraisal is deflationary, and does not involve heavy ontological commitments. Emotions count as normatively appraising their objects, on their view, simply in virtue of the way they affectively and motivationally engage with their

[15] Not all defenders of motivational theories of emotions deny that emotions have any normative content. See e.g. Scarantino (2014), who adapts Prinz's (2004) teleosemantic account of emotional representation to a motivational theory of emotion.

objects. In terms of its ontological commitments, D'Arms and Jacobson's account of emotional appraisal is on a similar footing to the accounts of normative beliefs defended by quasi-realist expressivists.[16] So, D'Arms and Jacobson do, it seems, deny that emotions have robust, ontologically committing normative content. Moreover, it is a central part of their view that the normative appraisals involved in emotions are ultimately 'response-dependent': they are rough-and-ready glosses that cannot in the end be spelled out without making reference to emotional syndromes (2022: 18-19, 2023: 136-152). Hence, the deflationary sense in which D'Arms and Jacobson claim that emotions appraise their objects does not lead to circularity when combined with neo-sentimentalist analyses.

The second distinctive feature of D'Arms and Jacobson's theory of emotions is its limited range of application. It is meant to apply only to what they call 'natural emotions', which are 'pan-cultural psychological kinds', and which encompass fear, anger, guilt, shame, disgust, and several other emotions (2022: 16, 2023: 116-126). Now, D'Arms and Jacobson take resentment and indignation not to be natural emotions, so it might be thought that, whatever its general compatibility with neo-sentimentalist analyses, defenders of MB cannot appeal to their theory of emotion to avoid circularity objections. However, D'Arms and Jacobson distinguish two kinds of emotions besides natural emotions: what they call 'cognitive sharpenings' and 'motivational sharpenings' (2003: 137-138). 'Cognitive sharpenings' are emotions that are developed from natural emotions by involving distinctive thoughts. One example they give is homesickness, understood as a cognitive sharpening of sadness involving the thought that you have left home. 'Motivational sharpenings' are emotions that are developed from natural emotions by involving more specific kinds of motivational states. They

[16] See Sinclair (2021: 181-186) for a recent defence of a quasi-realist expressivist account of moral beliefs. D'Arms and Jacobson clearly indicate that they take their overall position in *Rational Sentimentalism* (including, presumably, their account of emotional appraisal) to be compatible with quasi-realism (2023: 4, 78). This lends further support to my claim that their account of emotional appraisal is deflationary, and does not involve heavy ontological commitments.

give vengefulness as a possible example, understanding it, presumably, as a motivational sharpening of anger involving motivation towards getting revenge.

Now, D'Arms and Jacobson themselves take resentment and indignation to be cognitive sharpenings of anger, involving deontic moral concepts (cf. 2003: 143). This account of resentment and indignation is not available to defenders of MB, at least insofar as we aspire to defend non-circular analyses. But D'Arms and Jacobson's account of resentment and indignation is an optional commitment. Consistently with their overall theory, we might instead develop accounts of these emotions on which they are motivational sharpenings. To do so, all we would need to do is define clearly their specific motivational goals and show that resentment and indignation, as understood in terms of these goals, are psychologically real phenomena. And, as we saw in Chapter 2, this can be done. These emotions can be characterised in terms of their goal of getting offenders to hold themselves accountable by feeling guilty (for the right reasons) and making amends. There is ample evidence that, so understood, resentment and indignation are psychologically real, and indeed the typical driving force behind overt moral blaming behaviour (Dill and Darwall 2014: 46-54). So, at first blush, it seems that there is room within D'Arms and Jacobson's overall motivational theory of the emotions for defending the claim that moral blaming emotions (including indignation) do not, at the deepest level, have any normative content whatsoever.[17]

Like the James-Lange theory, D'Arms and Jacobson's motivational theory of emotions claims that (at least some) emotions lack any normative content whatsoever, at the deepest level. But in contrast with the James-Lange theory, it is much better placed to explain why

[17] A second motivational theory of emotions according to which (at least some) emotions do not have any normative content whatsoever is Deonna and Teroni's 'attitudinalist' theory (2012, 2015, forthcoming). Deonna and Teroni themselves argue that their theory of emotions fits nicely with (non-circular) neo-sentimentalist analyses of normative concepts (2021). Since the difficulty with motivational theories I explain shortly arises equally for Deonna and Teroni's theory (in virtue of clear, central commitments), I do not discuss their theory at length in the main body of the text.

emotions are subject to fittingness assessment (cf. D'Arms and Jacobson 2023: 136-152). In general, motivational states are typically assessable in terms of fittingness and closely related normative categories such as reasons. For example, assessment of desires and intentions in these terms is very familiar. So, D'Arms and Jacobson's theory of emotions is a good fit for neo-sentimentalist analyses: it holds that (some) emotions do not, at the deepest level, have normative content, yet it explains why emotions are assessable in terms of fit.

However, there is an important challenge to D'Arms and Jacobson's motivational theory, and any motivational theory that resembles it in claiming that (at least some) emotions do not have any normative content whatsoever at the deepest level. The difficulty is that, for any such theory to be workable, two conditions need to be met, but meeting both of them jointly is challenging. The first condition is that the motivational states involved in emotions cannot have goals such that, for someone to be in that goal-directed motivational state, they must have a prior normative representation of some kind (cf. Roberts 2003: 157-176; Ballard 2021; Tappolet 2023: 83-92). As I will put it, the motivational goals of emotions cannot be 'normative content-implicating'. To appreciate this condition, suppose, for the sake of argument, that the motivational goal of fear is *protection from danger*. In order to be motivated to protect yourself, or something you care about, from danger, it seems that you must antecedently represent the relevant situation as dangerous (Tappolet 2023: 83). We could not intelligibly ascribe *this* motivation to someone without also ascribing this prior normative representation. Moreover, this normative representation would not be a 'response-dependent' appraisal in D'Arms and Jacobson's sense. That is, the claim that fear represents its objects as dangerous would not provide a rough-and-ready gloss that cannot in the end be spelled out without making reference to the emotional syndrome of fear. On the contrary, we would have to appeal to dangerousness to characterise the emotional syndrome of fear itself.

The second condition that must be met is a general constraint on any adequate theory of emotions: it must be able to distinguish different emotion-types successfully. For example, an adequate theory of emotions must enable us to explain the difference between fear and disgust. Any adequate motivational theory of emotions that claims that at least some emotions do not have any normative content whatsoever is committed to distinguishing the relevant emotion-types by their motivational goals, without appealing to motivational goals that are normative content-implicating. But this will not be easy to do. Fear and disgust, for example, both motivate us to avoid their objects, or protect ourselves from them. They motivate us to avoid their objects in different ways, of course, but a tempting way of spelling out this difference is to say that the motivational goal of fear is protection from danger, whereas the motivational goal of disgust is protection from contamination (Tappolet 2023: 83). However, these goals are normative content-implicating: they implicate prior normative representations of danger and contamination, respectively. We might instead try to explain the difference between the motivational goals of fear and disgust without appealing to motivational goals that are normative content-implicating. Perhaps, for instance, disgust motivates sensory avoidance in a way that fear does not (cf. D'Arms and Jacobson 2022: 19). But it remains to be seen how far it will be possible to discriminate adequately among the motivational goals of emotion-types without appealing to goals that implicate normative content.

The relevance of all of this to MB is as follows. If we want to argue that indignation has no normative content whatsoever while appealing to a motivational theory of emotions, we will need to characterise the motivational goal of indignation in a way that does not implicate an antecedent normative representation. The motivational goal of indignation, we saw, is to get offenders to hold themselves accountable by feeling guilty (for the right reasons) about what they have done with respect to some other person(s) (that is, some other person(s) than the indignant agent) or impersonal value and make amends in the appropriate way. The question,

then, is whether we can give a plausible interpretation of this motivational goal on which it is not normative content-implicating. Although I want to leave open the possibility that this can be done, there are several obstacles that would need to be overcome in order to provide such an interpretation.

One issue concerns how making amends is to be understood. Some prominent ways of understanding making amends seem to make the motivational goal of indignation normative content-implicating. For example, it is often claimed that making amends withdraws threatening messages that are typically sent out by unexcused violations of certain standards, such as the message that these standards are unimportant, or that the people who were principally disadvantaged by their violation are unimportant (cf. Murphy 1982: 509; Hieronymi 2001: 548-549; Griswold 2007: 55-56; Radzik 2009: 92-97). I will draw on this work and develop an account of making amends along these lines in 5.2.3. However, whatever their other virtues, accounts along these lines seem to make the motivational goal of indignation normative content-implicating: if making amends is understood in this way, then someone who is motivated to get someone to make amends for φ-ing must surely represent φ-ing as an action of a kind that typically sends out such threatening messages.

But perhaps we might understand making amends in a different way. For example, perhaps we could understand making amends in terms of relationship repair, and understand the relationships at issue as ones in which conflict and strife are kept to a minimum, and the conditions for mutually advantageous co-operation are secured.[18] This would seem not to make indignation normative content-implicating (at least given a suitable account of what counts as 'mutually advantageous' co-operation). But it still leaves the problem of impersonal values. People are sometimes indignant in response to actions not only, or even primarily, because of

[18] Allan Gibbard (1990: 67-68, 139-140, 146-150) seems to have some such conception of making amends in mind when he characterises moral blaming emotions.

how these actions affect other persons, but because of other effects of these actions: for example, their effects on the natural world. When indignant agents are motivated to get offenders to feel guilty (for the right reasons) and make amends for such actions, the amends at issue cannot plausibly be understood in terms of repairing relationships that keep strife and conflict to a minimum, and secure the conditions of mutually advantageous co-operation. But perhaps we might understand amends in such cases simply in terms of restoration: what indignant agents would then be motivated to do, in such cases, is to get offenders to feel guilty (for the right reasons) and undo the effects of their actions insofar as this is possible.

This is just a sketch of how we might try to interpret the motivational goal of indignation in such a way that it is not normative content-implicating. As I indicated above, I want to remain neutral on whether this strategy might prove successful. But my defence of MB is not hostage to the fortunes of this attempt. In the next two sections, I argue that MB can be defended as a non-circular conceptual and metaphysical analysis even if it is granted that indignation has normative content.

5.2.2 Normative Content and Redundancy

There is an important objection that needs to be overcome to show that non-circular neo-sentimentalist analyses of normative concepts and properties/relations, such as MB, can be defended even if it is granted that the emotions that feature in these analyses have independently intelligible normative content. By 'independently intelligible' normative content, I mean normative content that can be understood without reference to these emotions (unlike what D'Arms and Jacobson call 'appraisals'). The objection is due to D'Arms and Jacobson (see, e.g., 2003: 127-128, 2023: 91). I will first illustrate the objection with a simple dummy neo-sentimentalist analysis of the concept and property of contemptibleness:

> *Contemptible*: Something is contemptible iff (Def) it is fitting for anyone to feel contempt towards it.

The objection takes the form of a dilemma: if it is granted that contempt has independently intelligible normative content, Contemptible must be either circular or redundant. It is circular if contempt represents its objects as contemptible. And it is redundant if it has different normative content: for example, if it represents its objects as ranking low in terms of an important standard of personal assessment. For it seems that we might then analyse contemptibleness directly in terms of this normative content, without referring to the fittingness of contempt at all. That is, we might simply claim that something is contemptible iff (Def) it ranks low in terms of an important standard of personal assessment.

One way to respond to this dilemma would be to embrace circularity, but deny that the circularity is vicious. As I have already said, my aim is to defend MB as a non-circular analysis. Hence, I will aim to resist the second horn of the dilemma. The claim that neo-sentimentalist analyses are redundant if the emotion in question has normative content that is different from that which is the target of the analysis can be resisted if we reject EFAR in favour of HEF. Since we have good reasons to reject EFAR in favour of HEF anyway, this gives us an independently motivated response to the objection.

If we accept EFAR, and claim further that contempt represents its objects as ranking low in terms of an important standard of personal assessment, then Contemptible claims, in effect, that something is contemptible iff (Def) it would be accurately represented as ranking low in terms of an important standard of personal assessment. It does seem that we might then drop talk of accurate representation from the analysis without loss, and analyse contemptibleness directly in terms of the content of contempt. So, the second horn of D'Arms and Jacobson's dilemma seems compelling if we accept EFAR.

But if we reject EFAR in favour of HEF, this horn of the dilemma is much less compelling. Contempt tends to motivate social withdrawal: contemnors typically seek to distance themselves socially from those towards whom they feel contempt (Bell 2013: 45-46, 53-54; Fischer and Giner-Sorolla 2016: 347-348). This can involve such things as not being open to close personal relationships with these agents, such as friendship, and also withholding certain gestures of respect and honour (such as refusing to shake hands). According to HEF, the fittingness of contempt is a function of the fittingness of both its representational and motivational components. When we plug this account of the fittingness of contempt into Contemptible, we get the claim, in effect, that something is contemptible iff (Def) emotional representation of it as ranking low in terms of an important standard of personal assessment would be fitting, *and* emotional motivation that aims at social withdrawal from those who exemplify it would be fitting. Emotional fittingness does not simply drop out of *this* analysis as redundant, in favour of an analysis directly in terms of the normative content of contempt. There are plenty of reasons why we might think that Contemptible, when paired with HEF, provides a deeper and more informative analysis. In particular, it provides a better explanation of the social significance of the contemptible, and why we have good reasons to care about which personal characteristics are contemptible. (Note that this need not involve denying that directly analysing contemptibleness in terms of the normative content of contempt would be extensionally adequate; only that it would be explanatorily inadequate).

In the next section, I develop this general argumentative strategy at greater length with respect specifically to MB. I argue that, when paired with HEF, MB can give an informative, non-circular analysis of deontic moral concepts and relations even if indignation has independently intelligible (non-moral) normative content.

5.2.3 An Account of Making Amends

The proposal to pair MB with HEF in this way faces three problems that need to be overcome. First, it might be worried that it is open to a further circularity objection. The concern is that the motivational goal of indignation, which is, roughly, to get offenders to hold themselves accountable by feeling guilty (for the right reasons) and making amends, cannot be understood independently of moral wrongness. The worry is that we cannot give a satisfactory account of what is involved in making amends without making ineliminable reference to moral wrongness. Second, we need an argument for thinking that, even if indignation has normative content, it does not have deontic moral content. Third, pairing MB with HEF might be thought to give rise to a new concern about redundancy, to the effect that we could analyse moral wrongness directly in terms of the normative content of moral blaming emotions *and* the goals at which these emotions aim, without making essential reference to these emotions themselves. I address each of these challenges in turn.

To meet the first challenge, defenders of MB need to provide an account of making amends that characterises it independently of its connection to moral wrongness. Fortunately, we will not have to develop such an account from scratch: the literature on making amends already contains resources for doing this. Generally, an account of making amends needs to identify the harms that making amends can help to repair and explain how the various actions by which we make amends (apologies, offers of compensation, etc.) can help to repair those harms. Work on making amends has uncovered various kinds of reparative work that making amends can do. It is possible to characterise these kinds of repair without recourse to deontic moral concepts. One important kind of reparative work that making amends can do is to withdraw threatening messages that are sent out by φ-ing, such as the message that the standards φ-ing violated are unimportant and/or that certain agents who were adversely affected by φ-ing are unimportant (cf. Murphy 1982: 509; Hieronymi 2001: 548-549; Griswold 2007:

55-56; Radzik 2009: 92-97).[19] Apologies and further reparative gestures such as compensation can withdraw these messages by indicating that their author does not stand by them. A related kind of reparative work that making amends can do is to repair damaged trust in others to respect the violated standards going forward (Walker 2006: 191-229). Finally, making amends can serve to compensate for material harms, such as damage to property.

According to MB, standards of moral wrongness are standards such that violation of them without a moral excuse merits emotionally charged responses that seek to ensure that the offender repairs the damage wrought by their standard-violation in these ways (insofar as this is possible). This provides a non-circular explanation of what distinguishes moral wrongness from other kinds of impermissibility. Consider prudential impermissibility, for instance. Someone who claims that acting in ways that are deeply imprudent is not always morally wrong can on this approach be understood as claiming that deep imprudence does not always merit these responses. Prudentially impermissible actions harm their agents, send out the message that their interests are unimportant and/or that prudential standards generally are unimportant, and damage trust that the offender will act in prudent ways. To claim that deeply imprudent actions are not always morally wrong is to claim that these actions do not always merit emotionally charged, community-wide responses that aim at repairing these harms, withdrawing these messages, and restoring this damaged trust.

The second thing we need to do if we are to defend MB as a non-circular analysis by pairing it with HEF is to provide an argument for thinking that, even if indignation has normative content, it does not have deontic moral content. Now, we saw in 5.2.1 that understanding making amends on the lines suggested here renders the motivational goal of indignation normative content-implicating. Someone who is motivated to get someone to make

[19] The notion of importance at play here is generic importance, rather than specifically moral importance.

amends for φ-ing, on this understanding of what making amends involves, must surely represent φ-ing as an action of a kind that typically sends out the threatening messages that the standards φ-ing violated are unimportant and/or that certain agents who were adversely affected by φ-ing are unimportant. This is normative content, but it is not deontic moral content. However, it might be suggested that, in additional to representing this (non-moral) normative content, indignation also represents its objects as morally wrong.

There are good reasons for resisting this suggestion, however. As we saw in Chapter 2, resentment and guilt are often fitting in response to (and are often in fact felt in response to) actions that are not morally wrong. Many violations of standards of good friendship are like this, as when one friend does not make an effort to keep in touch with another. So, it is not plausible to think that resentment and guilt have deontic moral content. A friend who resents another for violating standards of good friendship need not represent their action as violating standards of moral wrongness. But then why think that indignation has deontic moral content? As I explained in Chapter 1, I use 'resentment' and 'indignation' to distinguish personal from vicarious responses. Resentment aims at making offenders feel guilty (for the right reasons) about what they have done to *you* and make amends with *you*. Indignation, on the other hand, aims at making offenders feel guilty (for the right reasons) about what they have done with respect to some other person(s) or impersonal value and make amends in the appropriate way. There is no need to introduce deontic moral content to indignation to provide a full characterisation of it. Indignation differs from resentment simply in that it is felt on behalf of another (or out of concern with an impersonal value).

The third objection to the claim that MB can fruitfully be paired with HEF is that this pairing renders moral blaming emotions redundant, since the analysis could be directly formulated in terms of their normative content and motivational goals. For instance, a simple analysis along these lines is that it is morally wrong for an agent to φ iff (Def) φ-ing violates

standards that typically send out threatening messages and it is fitting for that agent to make amends for φ-ing. This removes guilt, resentment, and indignation from the analysis, but since these emotions are characterised in terms of their links to making amends and their normative content it is not clear what is lost.

This attempt to eliminate guilt, resentment, and indignation should be resisted on the grounds that even though these emotions are characterised in terms of their connection to making amends, they are themselves crucial to the repair that making amends achieves. In particular, the reparative effects of apologies and compensation are due in large part to their being expressions of the offender's guilt-feelings and their conviction that these feelings, along with resentment and indignation, are fitting.

We saw earlier that two important kinds of moral repair that making amends can achieve are, first, to withdraw threatening messages sent out by φ-ing, and, second, to repair damaged trust. Painful feelings of guilt about φ-ing indicate that the offender cares about the standards they violated in φ-ing. Because of this, expressing these feelings and acknowledging their fittingness by apologising and offering compensation withdraws the message that the standards φ-ing violated are unimportant. Guilt-feelings are also relevant to withdrawing the message that certain agents who were adversely affected by φ-ing are unimportant. Insofar as guilt involves an intrinsic desire to make amends with such agents, feelings of guilt and their expression in apologies and compensation withdraw the message that these agents are unimportant and instead convey the message that they are worthy of respect. Acknowledging that resentment and indignation on their part are fitting also contributes to this. Finally, insofar as guilt is a source of motivation to avoid future standard-violations, guilt-feelings are important to repairing damaged trust.

5.3 Conclusion

This chapter examined various circularity objections to MB. I argued that all of these objections can be met. I argued, moreover, that meeting these objections does not require us to deny that indignation, the central moral blaming emotion, has any normative content – only that it has deontic moral content. In being compatible with the view that indignation has (non-moral) normative content, this defence of MB against circularity objections allows us to remain largely neutral on controversial issues in the philosophy of emotion.

Chapter 6

Further Non-Extensional Objections

This chapter considers and rejects further non-extensional objections to MB besides circularity objections, which were discussed in Chapter 5. To recap, MB holds:

> *Moral Wrongness as Moral Blameworthiness (MB)*: It is morally wrong for an agent to φ iff (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, where 'φ' stands for an object of deontic moral assessment.

As we saw in Chapter 1, MB can also be formulated in terms of moral obligation/requirement and moral permissibility. To ease exposition, the third challenge I consider is framed in terms of moral requirement.

I consider three different challenges to MB. All of the objections I consider primarily target MB as a conceptual analysis. Accordingly, Section 6.1 includes a brief recapitulation of my understanding of conceptual analysis, which was laid out more fully in the Introduction. Although this chapter addresses objections, the discussion is not purely defensive: seeing why the third objection is unsuccessful (in 6.3) reveals a further strength of MB.

Section 6.1 considers the objection that those whom I will call guilt and/or resentment and indignation 'rejecters' – people who claim that guilt and/or resentment and indignation would rarely or never be fitting – pose problems for MB, insofar as rejecters might seem able to make deontic moral judgments without contradicting themselves or suffering from conceptual confusion. I argue that insofar as the kind of conceptual analysis I defend is partly revisionary, rejecters do not pose problems for it.

Section 6.2 discusses the objection that MB is undermined by a thought experiment known as 'Conative Moral Twin Earth'. In this thought experiment, two separate communities of speakers use 'morally wrong' in very similar ways – in particular, this term plays very similar roles in deliberation and criticism. However, whereas the first community's use of 'morally wrong' is robustly linked to guilt, resentment, and indignation, the second community's use of this term is robustly linked to different emotions, such as shame and contempt. Insofar as we have the intuition that speakers from the first community can engage in substantive disagreement with speakers from the second by using 'morally wrong', this might be taken to suggest that MB does not provide the correct analysis of MORALLY WRONG. Against this, I argue, first, that taking a fine-grained look at the critical role played by MORALLY WRONG shows that it could not play this role if it was robustly linked to emotions other than guilt, resentment, and indignation; and, second, that substantive disagreement does not require that each community uses 'morally wrong' to express the same concept.

Section 6.3 discusses the objection that MB cannot adequately account for the moral praiseworthiness of actions performed from the motive of moral obligation. I argue that, given plausible assumptions about the connection between moral blame and making amends, and the significance of making amends for respecting persons, MB can adequately account for the moral praiseworthiness of such actions.

6.1 Guilt and/or Resentment and Indignation Rejecters

MB analyses MORAL WRONGNESS partly in terms of MORAL BLAMEWORTHINESS. We saw in Chapter 2 that an agent is morally blameworthy for φ-ing just in case they would be a fitting target of moral blaming emotions for φ-ing – guilt, resentment, and, especially, indignation. Now, some people hold that moral blaming emotions would rarely or never be fitting. For

example, Glen Pettigrove and Koji Tanaka argue that there are important strands in the history of both Eastern and Western philosophical thought according to which anger (of which resentment and indignation are forms) would rarely or never be fitting (2014). And some philosophers have taken dim views about the fittingness of guilt. For example, Gilbert Harman holds that 'guilt is not reasonable, appropriate, or warranted for people who have adequate motivation to act morally without being susceptible to guilt feelings' (2009: 211).

Let me call thinkers who hold that guilt and/or resentment and indignation would rarely or never be fitting 'rejecters'. Now, it might seem that rejecters pose difficulties for MB. At least some (apparent) rejecters seem to make moral wrongness judgments despite their rejection of the fittingness of guilt and/or resentment and indignation. Defenders of MB, it might be thought, are committed to claiming that such rejecters either (A) do not really make moral wrongness judgments, or else (B) contradict themselves or suffer from conceptual confusion. But this might be found implausibly uncharitable – surely it is not credible to exclude so many careful and reflective thinkers from making moral wrongness judgments, or to convict them of these kinds of mistakes.

To assess this objection, I will need to recapitulate some points about the kind of conceptual analysis I defend. In the Introduction, I argued that a fruitful style of conceptual analysis is that which aims to meet both a descriptive desideratum and a prescriptive desideratum. The descriptive desideratum states that the analysis must be broadly in line with the concepts we in fact use. More precisely, the analysis had better, for the most part, respect the platitudes associated with the relevant concept, where 'platitudes' is a term of art for statements of the inferential and judgmental dispositions that are typically had by those who possess this concept. The prescriptive desideratum states that the concept, as understood by the analysis, had better be a valuable concept that is worth keeping. In Chapter 3, I argued that the account of deontic moral concepts provided by MB explains how deontic moral concepts serve

two valuable social functions: 'the Deliberative Reliability Function' and 'the Respect Function'. In calling MB partly prescriptive, I mean that it is to some extent a proposal for how we *should* understand deontic moral concepts.

Given this understanding of conceptual analysis, the commitments of defenders of MB with respect to rejecters are as follows. Either (A) rejecters do not really make moral wrongness judgments *that deploy deontic moral concepts as they should be understood*, or else (B) rejecters contradict themselves or suffer from conceptual confusion. Now, it is not clear why these commitments create any problems for defenders of MB. In particular, it is not clear that these commitments must lead defenders of MB to make sufficiently uncharitable claims about rejecters so as to call MB into question. Mistakes about how deontic moral concepts should be understood, where this is partly a question of what concepts it would be valuable for us to have, are mistakes that even highly reflective and careful thinkers can make. Of course, defenders of MB need to show that it really does meet the descriptive and prescriptive desiderata on successful conceptual analyses. Chapters 2 and 3 were devoted to just this task. However successful these arguments were, it is hard to see how rejecters could pose an independent challenge to MB.[1]

Rather than simply concluding here, it is worth considering the views of some (apparent) rejecters in more detail. This will accomplish two things. First, it will show that genuine rejecters may be less common than is often thought. Second, it will allow us to see in more detail the kinds of mistakes that MB imputes to rejecters who seemingly go on to make moral wrongness judgments.

It is striking that (apparent) guilt and/or resentment and indignation rejecters typically do not reject the platitudes associated with MORALLY WRONG on which MB focuses. My

[1] One possibility is that we could construct 'Moral Twin Earth' cases involving rejecters. I discuss Moral Twin Earth objections to MB in the next section.

defence of MB especially emphasises the connection between moral wrongness and making amends. 'If it is morally wrong for an agent to φ, then, *ceteris paribus*, there is a reason for that agent to make amends for φ-ing without a moral excuse' captures an inferential disposition that is central to mastery of MORALLY WRONG. As we saw in detail in 2.2.1, MB can respect this platitude by appealing to the motivational tendencies of moral blaming emotions: guilt motivates agents to make amends, and resentment and indignation motivate agents to get offenders to feel guilty and make amends.

In the course of arguing that many thinkers in Eastern and Western philosophical traditions who reject the fittingness of anger nonetheless possess MORALLY WRONG, Pettigrove and Tanaka emphasise that such thinkers 'can distinguish… …between actions that require expiation and those that do not' (2014: 276). And when Harman sketches a model of an admirable moral agent who is not susceptible to guilt feelings, he writes: 'the admirable people I have in mind feel regret about moral mistakes, but not guilt… …they can apologize, say that they are sorry for what they have done, try to make amends, and sincerely promise not to do it again' (2009: 211). As we will see, the fact that these thinkers do not reject the platitude linking moral wrongness with making amends in fact casts some doubt on whether they are correctly interpreted as rejecters at all.

When Pettigrove and Tanaka explain how the figures they discuss[2] understand anger, they characterise the phenomenology and motivational tendencies of this emotional response as follows:

> …at the heart of the experience of anger are the desire to lash out and the desire that the object of one's anger suffer pain, verbal abuse, or disgrace (preferably at one's own

[2] They focus chiefly on Śāntideva and Cassian, but they take these two thinkers to be representative of many thinkers in Eastern and Western philosophical traditions who have argued that anger would rarely or never be fitting (275).

> hands)… …What each of [Śāntideva and Cassian] calls anger commonly manifests itself in retaliatory actions. (2014: 275)

The emotion that Pettigrove and Tanaka describe here is importantly different from resentment and indignation. The motivational goals of resentment and indignation are to make offenders hold themselves accountable by feeling guilty (for the right reasons) and making amends for their conduct, not to retaliate (Dill and Darwall 2014: 46-54). Hence, the fact that many thinkers in Eastern and Western philosophical traditions reject the fittingness of an emotion that is characterised in terms of the phenomenological and motivational features that Pettigrove and Tanaka describe does not provide evidence that these thinkers reject the fittingness of resentment and indignation. Moreover, insofar as these thinkers acknowledge that there are actions that 'require expiation' – as they must, if it is to be plausible to interpret them as possessing and deploying MORALLY WRONG – it is not obvious that they *would* reject the fittingness of resentment and indignation.

Similar points apply to Harman's discussion of guilt. Harman explains that the conception of guilt-feelings he targets 'identifies them with feelings of remorse, involving deep regret, painful humiliation, distress, self-punishment, and/or self-flagellation' (2009: 205). This is importantly different from the understanding of guilt-feelings that I work with here. Although guilt, on my understanding, is a painful emotion, it has little or no connection with motivation towards self-punishment or self-flagellation. Instead, it is distinctive in virtue of its motivational goal of making amends (Tangney and Dearing 2002: 19; De Hooge 2019). The distinctive pain of guilt is the pain that comes with being conscious of non-satisfaction of its goal of making amends, and hence, typically, having damaged relationships that are important to you. It is not clear that Harman would reject the fittingness of *this* emotion, given his remark quoted above that admirable moral agents feel regret about moral mistakes, and are capable of

apologising and otherwise making amends (2009: 211). Indeed, the 'regret' that Harman speaks of in this passage may just be guilt, as MB understands it.

There is, however, another interpretation of Harman's views, and considering this interpretation will allow us to see the kinds of mistakes that MB might impute to rejecters who seemingly go on to make moral wrongness judgments. On this interpretation, Harman rejects the fittingness of guilt even as MB understands it, but aims to respect the platitude linking moral wrongness with making amends by holding that there are ways of making amends that do not involve expressing guilt-feelings. Consider in this connection the following passage: 'morally good people with no disposition to feel guilt will not want to benefit from any wrongful acts and so will try to make amends in some other way than by feeling or pretending to feel guilt' (2009: 211).

What mistake does MB impute to a rejecter such as Harman, on this interpretation of his views? We saw in Chapter 5 that guilt, resentment, and indignation, together with the judgments that these emotions are fitting, are crucial to the repair that making amends achieves. In particular, the reparative effects of apologies and compensation are due in large part to their being expressions of the offender's guilt-feelings and their conviction that these feelings, along with resentment and indignation, are fitting. Two important kinds of moral repair that making amends can achieve are, first, to withdraw threatening messages sent out by φ-ing, and, second, to repair damaged trust. Painful feelings of guilt about φ-ing indicate that the offender cares about the standards they violated in φ-ing. Because of this, expressing these feelings and acknowledging their fittingness by apologising and offering compensation withdraws the message that the standards φ-ing violated are unimportant. Guilt-feelings are also relevant to withdrawing the message that certain agents who were adversely affected by φ-ing are unimportant. Insofar as guilt involves an intrinsic motivation to make amends with such agents, feelings of guilt and their expression in apologies and compensation withdraw the message that

these agents are unimportant and instead convey the message that they are worthy of respect. Acknowledging that resentment and indignation on their part are fitting also contributes to this. Finally, insofar as guilt is a source of motivation to avoid future standard-violations, guilt-feelings are important to repairing damaged trust. Moreover, insofar as our emotions are to a significant extent responsive to our judgments as to their fittingness, guilt-feelings that are backed by fittingness judgments are more reliable than guilt-feelings which are not backed up by such judgments, and are hence better able to repair damaged trust.

The mistake that MB imputes to a rejecter such as Harman is to overlook the importance of guilt-feelings to making amends. Insofar as an analysis of MORALLY WRONG aims to respect the link between moral wrongness and making amends – as it should, since this link is important to the value of this concept – it should be formulated in terms of the fittingness of moral blaming emotions, including guilt.

6.2 Conative Moral Twin Earth

The next objection I will discuss is a development of the 'Moral Twin Earth' objection, which was originally deployed against versions of moral realism that apply a causal theory of reference to moral terms.[3] As applied to MB, the objection begins with the following thought experiment:

> [Conative Moral Twin Earth] features an alien society that is much like our own. Their word 'wrong' functions much as our word 'wrong' does: as a term of criticism that plays a distinctive role in deliberation and social coordination. The term is used to characterize a very similar range of actions, and there are very similar controversies

[3] See McPherson (2013) for an overview.

> about its proper extension. The crucial difference is that the twins' use of 'wrong' seems to be robustly linked to emotions that we recognise to be contempt and shame rather than to resentment and guilt. (Björnsson and McPherson 2014: 8)

Before laying out the objection from Conative Moral Twin Earth, it is worth making one amendment to this thought experiment. Rather than simply characterising 'wrong' (or, better, 'morally wrong') on Twin Earth as a term of criticism, it is worth adding that it plays a very similar critical role to that played by 'morally wrong' on Earth. The reason for this amendment is that if 'morally wrong' played a different kind of critical role on Twin Earth than on Earth, this would be *prima facie* grounds for denying that it expresses the same concept as the Earth term 'morally wrong'. Hence, the thought experiment will be most effective if we assume that 'morally wrong' on Twin Earth plays a very similar critical role to 'morally wrong' on Earth. (I will consider at the end of this section whether the objection might do without this assumption). Starting from this thought experiment, the objection to MB runs as follows (cf. Merli 2008: 35; Björnsson and McPherson 2014: 9):

> P1: When an inhabitant of Earth applies 'morally wrong' to a certain act-type, A, and an inhabitant of Twin Earth applies 'not morally wrong' to A, the Earth inhabitant is in substantive disagreement with his or her twin.[4]
>
> P2: The Earth inhabitant is in substantive disagreement with his or her twin only if both use 'morally wrong' to express the same concept.
>
> P3 (from P1 and P2): The Earth inhabitant and his or her twin both use 'morally wrong' to express the same concept.

[4] 'Substantive' disagreement contrasts with merely verbal disagreement, as when I say the bank is on the left (meaning the river bank) and you say the bank is on the right (meaning the place where the money is). For more on substantive disagreement vs merely verbal disagreement, see Plunkett and Sundell (2013).

P4: MB does not provide the correct analysis of the concept expressed by uses of 'morally wrong' by inhabitants of Twin Earth.

P5: If MB does not provide the correct analysis of the concept expressed by uses of 'morally wrong' by inhabitants of Twin Earth, but *does* provide the correct analysis of the concept expressed by uses of 'morally wrong' by inhabitants of Earth, then the Earth inhabitant and his or her twin do not both use 'morally wrong' to express the same concept.

Therefore, C1 (from P3, P4, and P5): MB does not provide the correct analysis of the concept expressed by uses of 'morally wrong' by inhabitants of Earth.

I will make two objections to this argument. First, I will object to the thought experiment on which it is based. It assumes that two societies could use terms that play very similar roles in deliberation, social coordination, and criticism, but whereas one of these terms is robustly linked to guilt, resentment, and indignation, the other is robustly linked to different reactions such as shame and contempt. I will argue that taking a fine-grained look at the critical role played by MORALLY WRONG throws this assumption into doubt. Second, I will argue that P2 is false. The Earth inhabitant and his or her twin could substantively disagree with each other even if their uses of 'morally wrong' express different concepts, so long as they disagree with each other about *what to do*.

First, however, I will consider whether there is an easier way for defenders of MB to avoid the objection from Conative Moral Twin Earth. It might be wondered whether this objection passes my defence of MB by, insofar as I claim that MB is partly revisionary or prescriptive. If MB is at bottom a proposal concerning how we *should* understand deontic moral concepts, rather than a descriptive analysis of our *actual* deontic moral concepts, then (it might be thought) it does not need to account for our intuitions concerning Conative Moral Twin

Earth, since these intuitions concern our actual deontic moral concepts. However, the problem with this response is that MB is only partly prescriptive. MB, as understood here, is meant to align significantly with our actual deontic moral concepts. The revisionary or prescriptive aspect of MB is limited, insofar as MB is meant to meet the descriptive desideratum on successful conceptual analyses as well as the prescriptive desideratum.

In order for the objection from Conative Moral Twin Earth to get off the ground, it has to be possible for there to be a community of speakers whose use of the term 'morally wrong', despite being robustly linked to shame and contempt rather than guilt, resentment, and indignation, is nonetheless highly similar to our use of 'morally wrong'. However, when we look closely at the critical role played by MORALLY WRONG, there are compelling grounds for thinking this is not possible. In the last section, we saw that a central platitude surrounding MORALLY WRONG concerns the connection between moral wrongness and making amends. 'If it is morally wrong for an agent to φ, then, *ceteris paribus*, there is a reason for that agent to make amends for φ-ing without a moral excuse' captures an inferential disposition that is central to mastery of MORALLY WRONG. Hence, the critical role played by MORALLY WRONG is distinctive at least partly in terms of its link with making amends: agents who act morally wrongly in φ-ing are open to a kind of criticism such that there is usually a reason for them to make amends for φ-ing.

There are good reasons for thinking that a term that is robustly linked to shame and contempt, rather than guilt, resentment, and indignation, could not play this critical role. The motivational tendencies of shame are complex. It is associated in the short-term with motivation towards hiding or concealing the object of shame (Williams 1993: 90; Deonna et al. 2011: 174-179). The objects of shame can be various: someone might be ashamed of a character trait such as dishonesty, a different kind of personal shortcoming such as lack of wit, or a feature of their appearance, such as being short. It has been suggested that shame may be

associated in the long-term with motivation towards self-improvement or self-reform (Williams 1993: 90; Deonna et al. 2011: 174-179). While attempts at making amends may form part of a project of self-improvement, shame is much less tightly connected with making amends than guilt. Contempt tends to motivate avoidance and withdrawal: contemnors seek to distance themselves socially from the objects of their contempt (Bell 2013: 45-46, 53-54; Fischer and Giner-Sorolla 2016: 347-348). This is in stark contrast with the motivational goals of resentment and indignation, which are linked with pressuring offenders towards feeling guilty and making amends.

Reflection on the motivational goals of shame and contempt suggests that a term that is robustly linked with them could not play the critical role played by our term 'morally wrong'. Instead of imputing a criticism such that it is usually appropriate for agents who are open to this criticism to make amends, it would seem more likely to make a criticism such that it is usually appropriate for agents who are open to it to make efforts towards self-improvement or self-reform, and for others to distance themselves socially from them (for example, by being less willing to enter close relationships such as friendship with them). This suggests that the Conative Moral Twin Earth thought experiment sketched above may not be coherent: there could not be a term that played highly similar roles in deliberation, criticism, and coordination to our term 'morally wrong', but which was robustly linked with shame and contempt rather than guilt, resentment, and indignation.[5]

I will now develop a second response to the objection from Conative Moral Twin Earth. This response targets P2 (the Earth inhabitant is in substantive disagreement with his or her

[5] For further discussion of possible differences in ethical thought and practice between cultures that place greater emphasis on shame rather than guilt (and *vice versa*), see Morris (1976) and Creighton (1990). The argument of this paragraph generalises to the other kinds of blame besides moral blame that were distinguished in 2.1.2. Disgust, like contempt, is linked with avoidance (Giner-Sorolla et al. 2018). And the kinds of relationship modifications that Scanlon emphasises, such as adjusting your intentions and other attitudes towards someone, are not tightly linked with making amends either.

twin only if both use 'morally wrong' to express the same concept).[6] While substantive disagreement concerning whether a given act-type is morally wrong may require that the Earth inhabitant and his or her twin both use 'morally wrong' to express the same concept, this is not required for all kinds of substantive disagreement. In particular, insofar as both typically use conclusions about what is 'morally wrong' to guide them in their decisions about what to do, they could still be in substantive disagreement about what to do (cf. Copp 2000: 120-124; Merli 2002: 231-239). Drawing on this point, we could explain the intuition that the Earth inhabitant is in substantive disagreement with his or her twin while rejecting the claim that both use 'morally wrong' to express the same concept.

Note that this way of explaining intuitions of disagreement while denying concept identity could be used to complement the first response to the objection from Conative Moral Twin Earth developed above. For imagine that someone granted that there could not be a term that played *highly* similar roles in deliberation, criticism, and coordination to our term 'morally wrong', but which was robustly linked with shame and contempt rather than guilt, resentment, and indignation, but then went on to claim that it would suffice for disagreement that the Twin's term played *somewhat* similar deliberative, critical, and coordinative roles. They might then go on (*via* the argumentative sequence outlined above) to argue from the possibility of disagreement to the conclusion that MB does not provide the correct analysis of the concept expressed by our term 'morally wrong'. To counter this argument, we might explain the disagreement in question as a disagreement about what to do (this disagreement would parallel the way in which an inhabitant of a culture that prizes honour might, by characterising an action as honourable, disagree about what to do with someone who characterises it as morally wrong). This explanation would be particularly attractive, because the fact that our term 'morally wrong'

[6] This response takes inspiration from work by David Copp and David Merli criticising applications of Moral Twin Earth objections to various theories of the meta-semantics of moral discourse developed by naturalist moral realists (Copp 2000: 120-124, Merli 2002: 231-239). See also Plunkett and Sundell (2013).

plays an importantly different critical role from the Twin's 'morally wrong' already provides grounds for thinking that these terms express different concepts.

6.3 Moral Worth

Let me now consider a different objection to MB. It has been argued that MB runs into difficulties with respect to the issue of moral motivation (Owens 2012: 72-73; Tadros 2016: 25). The strongest version of this objection arises from the tension between MB and the following two claims:

> *The Motive of Moral Obligation is Sometimes Morally Praiseworthy* (*SMP*): Acting from a *de dicto*, intrinsic desire to act as you are morally required to act is sometimes (because of this) morally praiseworthy.
>
> *The Motive of Avoiding Moral Blameworthiness is Never Morally Praiseworthy (NMP)*: Acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is never (because of this) morally praiseworthy.

Let me make some clarifications. First, by 'moral blameworthiness-related standards', I mean standards such that, if an agent violated them without a moral excuse, they would be morally blameworthy for violating them. Second, the phrase in parentheses ('because of this') indicates that the desire in question contributes towards, and helps to explain, the moral praiseworthiness of the action it motivates. Third, by a '*de dicto* desire', I mean a *de dicto* report of a desire: this is a report of a desire that expresses the concepts that feature in the content of the desire. Hence, a *de dicto* desire to act as you are morally required to act is a desire that involves the concept MORALLY REQUIRED as part of its content. In other words, it is a desire to fulfil your moral requirements *under that description*. Fourth, by an intrinsic desire to φ, I mean a desire to φ for its own sake. With these clarifications in hand, we can see that defenders of MB must reject

either SMP or NMP (but not both). For if MB is correct, then to act from a *de dicto*, intrinsic desire to act as you are morally required to act just is to act from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards. I will argue for SMP and against NMP. I will start by explaining why both SMP and NMP are *prima facie* plausible, before objecting to NMP by arguing that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards can be morally praiseworthy insofar as, and because, it expresses respect for persons.

To begin with, it is worth distinguishing SMP (acting from a *de dicto*, intrinsic desire to act as you are morally required to act is sometimes (because of this) morally praiseworthy) from some other claims in the neighbourhood. In particular, SMP needs to be distinguished from:

> *Necessary Condition*: Being motivated by a *de dicto*, intrinsic desire to act as you are morally required to act is a necessary condition for an action to be morally praiseworthy.[7]

And:

> *Sufficient Condition*: Being motivated by a *de dicto*, intrinsic desire to act as you are morally required to act is a sufficient condition for an action to be morally praiseworthy.

SMP is different from, and does not entail, Necessary Condition. SMP holds that being motivated by a *de dicto*, intrinsic desire to act as you are morally required to act is one way an action can get to be morally praiseworthy, but it is not committed to the view that this is the only way this can happen. It is consistent with pluralist views of moral praiseworthiness for

[7] Strictly speaking, I take it that it is *agents* who are morally praiseworthy *for* actions, but I will sometimes speak of actions being morally praiseworthy as a *façon de parler*.

actions, according to which there is a plurality of different motivations in virtue of which actions can be morally praiseworthy (Isserow 2020).

Moreover, SMP is different from, and does not entail, Sufficient Condition. SMP holds only that it is *sometimes* the case that acting from a *de dicto*, intrinsic desire to act as you are morally required to act is (because of this) morally praiseworthy. This is consistent with thinking that further conditions need to be met if this desire is to result in morally praiseworthy actions. I will not attempt an exhaustive review of the possibilities here, but it is worth mentioning a couple of salient ones. First, it is plausible that some kind of epistemic condition (or conditions) need to be met. For example, Paulina Sliwa argues that, as well as acting from a desire with moral content, the agent must act from moral knowledge if they are to be morally praiseworthy for their action (2016). This view endorses SMP but rejects Sufficient Condition. It rules out, plausibly, that an agent could be morally praiseworthy for performing an action even if they act from grossly incorrect or unjustified moral views. Second, it may be that further conative states are needed if an action is to be morally praiseworthy. Perhaps, for instance, to be morally praiseworthy (or, at least, fully morally praiseworthy) an action needs to be motivated *both* by a *de dicto*, intrinsic desire to act as you are morally required to act *and* more direct desires concerning the reasons in virtue of which you are so required (such as, for example, the *de dicto*, intrinsic desire for Lisa's pain to stop). Again, this kind of view endorses SMP but rejects Sufficient Condition.

SMP is *prima facie* plausible, especially once it is distinguished from Necessary Condition and Sufficient Condition. All it claims is that being motivated by a *de dicto*, intrinsic desire to act as you are morally required to act is one factor that can contribute to moral praiseworthiness, even if further conditions need to be met in order for it to make this contribution. An example may help to bring out SMP's *prima facie* plausibility. Suppose that while Charlie is on his way to work he stops to assist an injured cyclist, Dean. Charlie has a *de*

dicto, intrinsic desire to act as he is morally required to act, and this desire, together with his belief that he is morally required to assist Dean, motivates him to do so. Suppose further that Charlie knows that he is morally required to assist Dean, that he understands the reasons why he is under this moral requirement, and that he is motivated additionally by more direct desires concerning these reasons. It seems plausible that Charlie is morally praiseworthy for stopping to assist Dean. Moreover, it seems plausible that being motivated by a *de dicto*, intrinsic desire to act as he is morally required to act contributes towards his moral praiseworthiness. It displays an admirable conscientiousness (cf. Isserow 2020: 542-550). In making these brief remarks, I do not mean to have firmly established SMP – that would take considerably more argument. But reflection on examples such as the one given above suffice to show that SMP is a *prima facie* plausible claim.

There is a well-known objection to SMP that is worth discussing briefly. The objection is due to Michael Smith (1994). Smith objects that motivation by *de dicto*, intrinsic desires to act as you are morally required to act involves a kind of fetishistic attachment to morality (75). Such motivation 'alienates [people] from the ends at which morality properly aims' (76). 'Good people', Smith writes, 'care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality, and the like' (75). However, it is not clear that this objection calls SMP into question so much as *Sufficient Condition*. It highlights the need for further conditions to be met if motivation by *de dicto*, intrinsic desires to act as you are morally required to act is to confer moral praiseworthiness (cf. Copp 1997b: 49-50, Isserow 2020: 549-550). Provided that people are aware of, and care non-derivatively about, the reasons why certain actions are morally required, it is hard to see how adding a *de dicto*, intrinsic desire to act as they are morally required to act could alienate them from the ends at which morality properly aims. Again, there is more to be

said about Smith's fetishism objection, and I do not take these brief remarks to show decisively that it fails. But they do help to establish the *prima facie* plausibility of SMP.

Let me turn now to NMP (acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is never (because of this) morally praiseworthy). NMP is also *prima facie* plausible. It will help to start by distinguishing the desire that NMP focusses on from another desire in this region. Suppose that Eric keeps his promise to Paula because he believes that Paula will morally blame him if he does not, and he wants not to be morally blamed because he finds moral blame unpleasant. Even though motivation of this kind may be common, it seems clear that an action that is motivated in this way is not morally praiseworthy. Being motivated by a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is importantly different from Eric's motivation, however. To be motivated by such a desire is not to be motivated to avoid *actually* being morally blamed, but rather to be motivated to avoid being such that one *would be a fitting target* of moral blame if one performed the relevant action without a moral excuse. Such motivation is not as *obviously* inapt for moral praise as Eric's motivation. Even so, it may seem unclear that it confers moral praiseworthiness on actions (cf. Owens 2012: 73). One way to bring this out is to see that it is not clear that such motivation would add anything to being motivated by more direct desires concerning the reasons in virtue of which moral requirements obtain. Suppose that Angela is motivated to help Faruq by a *de dicto*, intrinsic desire for his pain to stop. Now suppose that we add that Angela is also motivated by a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards. Does motivation by this second desire confer *additional* moral praiseworthiness on Angela's action? It is not clear that it does. NMP, then, is *prima facie* plausible.

Both SMP and NMP are *prima facie* plausible. Defenders of MB are committed to rejecting one of these claims (but not both). In the remainder of this section, I argue against

NMP. I argue that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is sometimes (because of this) morally praiseworthy. Given that SMP is, independently of MB, a *prima facie* plausible claim, the fact that MB dovetails with SMP is an additional strength of MB. Moreover, MB lends additional plausibility to SMP. This is because MB, together with the fact that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is sometimes (because of this) morally praiseworthy, helps to support and explain SMP. My argument that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is sometimes (because of this) morally praiseworthy will hinge on the connection between making amends and respect for persons. This connection was laid out at length in Chapter 3. I will summarise it more briefly here.

It seems clear that making amends for φ-ing can be an expression of respect – in particular, but not exclusively, for the person *with whom* you make amends through directed apologies, compensation, and the like (Radzik 2007: 75-109; Jonker 2020: 30-32). But it is less clear *how* making amends for φ-ing expresses respect. As we saw in Chapter 3, there are several ways in which making amends for φ-ing expresses respect. One way is due to the following feature of sincere apologies: sincerely apologising for φ-ing requires communicating to the addressee of the apology that you had strong normative reasons not to φ. Moreover, to apologise sincerely for φ-ing, it is usually not enough to communicate that *some consideration or other* was a strong normative reason not to φ. Rather, sincere apologies typically require communicating that certain salient facts about your victim were strong normative reasons not to φ – for example, that φ-ing significantly harmed them, or that they did not consent to your φ-ing. This is one way in which making amends for φ-ing expresses respect: it communicates to your victim (and others) that they are normatively significant, in the sense that certain salient

facts about them, such as facts about their interests, are strong normative reasons to treat them in certain ways and not in others.

Another way in which making amends for φ-ing expresses respect is by involving submission to the authority of your victim to determine, within reasonable bounds, what counts as adequate amends (Walker 2006: 200-201). The qualification 'within reasonable bounds' is important, because it is of course possible for victims to demand too much (or not enough) in the way of amends from moral wrongdoers. But within this range, victims have discretion to determine what will count as adequate amends. As Margaret Urban Walker puts the point, moral wrongdoers '[relinquish] primary authority over… …the measure of satisfactory response' (2006: 200).

Finally, to make amends for φ-ing, it is not enough *only* to apologise, offer compensation, or attempt to re-form your character. Making amends also requires accepting the authority of your victim and, perhaps, members of the wider community, to demand these things of you (Darwall 2006a: 82-86; Walker 2006: 200). Suppose that someone apologises to their victim, offers compensation, and attempts to re-form their character, but regards all of these things only as pre-requisites of virtue, rather than things that their victim could authoritatively demand of them. We might imagine that, in response to the demands of their victim for apology, compensation, and re-form, they reply, 'I'll do all of those things – but not because you expect them of me. I'll do them because it'll make me a better person'. Such an agent would seem not to have adequately made amends. This points to another way in which making amends for φ-ing expresses respect: it expresses respect for the authority of your victim and, perhaps, members of the wider community, to demand such things as apology, compensation, and attempts at re-form from you.

I will now draw on this account of the connection between making amends and respect for persons to argue that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is sometimes (because of this) morally praiseworthy. In being motivated by such a desire, an agent is motivated to avoid violating standards such that, if they violated them without a moral excuse, they would be a fitting target of guilt, resentment, and indignation. Guilt, as we have seen, motivates agents to make amends; complementarily, resentment and indignation motivate agents to make offenders feel guilty and make amends. Insofar as making amends expresses respect for persons, it seems plausible that motivation by a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards expresses respect for persons as well. To be motivated by such a desire is to acknowledge that failure to conform to the relevant standard without a moral excuse would make fitting a suite of emotional responses that aim at getting you to make amends for it.

In arguing that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards can be morally praiseworthy insofar as, and because, it expresses respect for persons, I mean to defend a claim that exactly parallels SMP (acting from a *de dicto*, intrinsic desire to act as you are morally required to act is sometimes (because of this) morally praiseworthy). Hence, I am not arguing that an action is morally praiseworthy *only if* it is motivated by a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards. Nor am I arguing that such motivation is a *sufficient condition* for moral praiseworthiness. Plausibly, the agent will need to meet certain epistemic conditions if they are to be morally praiseworthy for their actions, and perhaps further conative conditions will need to be met as well. My claim is only that being motivated by a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is one factor that can contribute towards moral praiseworthiness, even if further conditions need to be met in order for it to make this contribution.

Earlier, we saw that NMP (acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is never (because of this) morally praiseworthy) is supported by the thought that it is not clear that such motivation adds anything to motivation by more direct desires concerning the reasons in virtue of which moral requirements obtain. It is worth seeing how my defence of the claim that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards is sometimes (because of this) morally praiseworthy gives us resources for responding to this concern. To return to our earlier example, suppose that Angela is motivated to help Faruq by a *de dicto*, intrinsic desire for his pain to stop. Now suppose that Angela is also motivated by a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards. Why does this further desire make her more morally praiseworthy?

What this second desire adds – the reason why it confers additional moral praiseworthiness on Angela's action – is an expression of respect for Faruq. Now, we saw that one way in which making amends expresses respect for persons is by communicating to your victim (and others) that they are normatively significant, in the sense that certain salient facts about them, such as facts about their interests, are strong normative reasons to treat them in certain ways and not in others. Given that Angela is already motivated to help Faruq by a *de dicto*, intrinsic desire for his pain to stop, this aspect of the connection between making amends and respect for persons may not add anything to her original desire. But we saw that making amends also expresses respect for persons' normative agency – in particular, it expresses respect for the authority of your victim and, perhaps, members of the wider community, to demand such things as apology, compensation, and attempts at re-form from you, as well as submission to the authority of your victim to determine, within reasonable bounds, what counts as adequate amends. This does add something to Angela's *de dicto*, intrinsic desire for Faruq's pain to stop – something that can explain why she accrues additional moral praiseworthiness

by being motivated by a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards.

Before concluding this section, let me consider an objection. I have argued that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards can be morally praiseworthy insofar as, and because, it expresses respect *for persons*. In Chapter 2, we saw that some people hold that actions can be morally wrong in virtue of impersonal disvalues. For example, perhaps it would be morally wrong to cut down a beautiful old redwood tree for fun, even if doing so would not harm any person or violate any of their rights (Scanlon 1998: 172). Clearly, my argument for the claim that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards can be morally praiseworthy will not apply to such cases. However, it is not clear that this is a significant problem. First, we saw in Chapter 3 that making amends may also express respect for impersonal values, such as the value of natural beauty. If this is right, then this may support a parallel argument for thinking that acting from a *de dicto*, intrinsic desire to avoid violating moral blameworthiness-related standards can be morally praiseworthy as an expression of respect for such values.

Moreover, even if there can be moral wrongs that depend on impersonal values, such wrongs are typically paired with moral wrongs to persons. For example, perhaps damaging the natural environment is morally wrong because the natural environment has intrinsic value. But damaging the natural environment often wrongfully harms the interests of persons as well. In cases such as these, the argument given above for the moral praiseworthiness of actions performed from *de dicto*, intrinsic desires to avoid violating moral blameworthiness-related standards goes through. Moreover, moral wrongs to persons make up the central part of morality. It would not be a significant cost for MB if it is forced to deny that acting from a *de dicto*, intrinsic desire to act as you are morally required to act is ever (because of this) morally praiseworthy in the case of impersonal moral wrongs.

6.4 Conclusion

This chapter discussed further non-extensional objections to MB besides circularity objections, which were discussed in Chapter 5. I argued that MB is not vulnerable to objections from the possibility of rejecting the fittingness of guilt and/or resentment and indignation, from the 'Conative Moral Twin Earth' thought experiment, or from the conditions of moral praiseworthiness. In connection with this last issue, I argued that MB in fact gains plausibility from reflection on the conditions of moral praiseworthiness, insofar as it dovetails with the independently plausible claim that acting from a *de dicto*, intrinsic desire to act as you are morally required to act is sometimes (because of this) morally praiseworthy.

Chapter 7

Moral Wrongness and Normative Reasons

Consider the following two cases:

> *Fugitive Son*: A woman's son has committed a serious crime. She could hide him from the police, but if she does another innocent man will be wrongly convicted for the crime and sent to prison. (Wolf 2015: 41)
>
> *Williams's Gauguin*: Gauguin abandons his wife and children in Paris to go to Tahiti, where he can dedicate himself to painting. He produces brilliant work, work he would not have been able to produce had he stayed in Paris, and finds deep satisfaction and meaning in his artistic pursuits. The consequences for his wife and children, however, are dire: without their main source of financial support, they face great hardship and poverty. (Williams 1981: 22-26)

Both of these cases put pressure on the Overridingness Thesis. This is the thesis that if it is morally wrong for an agent to φ, then that agent has decisive normative reasons not to φ ('φ' stands for an action or omission).[1] At least on some ways of filling in further details, it seems clear that the mother acts morally wrongly if she hides her son from the police, and Gauguin acts morally wrongly in abandoning his wife and children in Paris. Yet for both cases there seems to be a further question as to whether these agents are justified from a comprehensive normative perspective. If the mother and Gauguin are justified from this perspective – if they have sufficient normative reasons to act in these ways, even though they act morally wrongly – then the Overridingness Thesis is false.

[1] This thesis is sometimes called 'Moral Rationalism' (cf. Portmore 2011, 2021; Archer 2014). Since it is a contentious question what the relation between rationality and normative reasons is, I prefer to avoid this label. 'Moral Rationalism' would seem apt only if rationality consists in responding correctly to your normative reasons, but some philosophers deny this (e.g., Broome 2013).

The question of the relative importance of morality as compared with other normative domains, such as prudence, is intrinsically interesting. An additional reason for caring about the Overridingness Thesis concerns the role this thesis plays in some influential arguments against certain moral theories, including traditional (i.e., maximising, agent-neutral) act-consequentialism. Some philosophers argue against traditional act-consequentialism on the grounds that it conflicts with the Overridingness Thesis (Stroud 1998: 171, 182-184; Portmore 2011: 29-32). According to this objection, agents do not always have decisive normative reasons to maximise agent-neutral value. Hence, given the Overridingness Thesis, it cannot always be morally wrong for them to fail to do so. If the Overridingness Thesis is false, objections of this kind will need to be rethought.

This chapter considers and rejects what is perhaps the most important argument for the Overridingness Thesis.[2] It remains neutral on whether the Overridingness Thesis itself is correct or incorrect. The argument I reject starts from MB:

> *Moral Wrongness as Moral Blameworthiness (MB)*: It is morally wrong for an agent to φ iff (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, where 'φ' stands for an action or omission.

Some philosophers have tried to derive the Overridingness Thesis from MB and a further premise linking being morally blameworthy for φ-ing with having decisive normative reasons not to φ (Gibbard 1990: 299-300; Darwall 2006a: 97-99, 2006b: 292; Skorupski 2010: 295-

[2] A presupposition of the Overridingness Thesis is that good sense can be made of unsubscripted normative concepts. Subscripted normative concepts include MORALLY OUGHT and PRUDENTIAL REASON. These normative concepts concern what we ought to do or have a reason to do by the lights of a special normative domain. We might represent this with a subscript: OUGHT_{M} and REASON_{P}, where '$_{\text{M}}$' stands for morality and '$_{\text{P}}$' stands for prudence. It is a contentious question whether, in addition to subscripted normative concepts, there is good sense to be made of unsubscripted normative concepts, such as DECISIVE NORMATIVE REASONS. The Overridingness Thesis presupposes, controversially, that good sense can be made of unsubscripted normative concepts, but I will not call this presupposition into question. Copp (1997) rejects this presupposition. See Dorsey (2016: Chapter 1) for a response.

301; Portmore 2011: 38-51, 2021: 51-62). I argue that this argument fails. My argument targets the premise linking being morally blameworthy for φ-ing with having decisive normative reasons not to φ. I go on to defend a different account of the upshots of MB for the relation between moral wrongness and normative reasons. I argue that MB supports only a weaker thesis I call 'the Normativity Thesis'. According to the Normativity Thesis, if it is morally wrong for an agent to φ, then that agent has strong normative reasons not to φ. However, whether these reasons are decisive or even sufficient is left open by this thesis.

Section 1 contrasts the Overridingness Thesis with rival views concerning the relation between moral wrongness and normative reasons, and outlines the argument from MB to the Overridingness Thesis. Section 2 undermines a crucial premise of this argument linking being morally blameworthy for φ-ing with having decisive normative reasons not to φ. Finally, Section 3 argues that MB supports only the Normativity Thesis.

7.1 MB and the Overridingness Thesis

What is the relation between moral wrongness and normative reasons? The Overridingness Thesis is one answer to this question, but it is not the only answer. It can be contrasted with the following three theses:

> *The Sufficiency Thesis*: If it is morally wrong for an agent to φ, then that agent has sufficient normative reasons not to φ.
>
> *The Normativity Thesis*: If it is morally wrong for an agent to φ, then that agent has strong normative reasons not to φ.
>
> *The No-Entailment Thesis*: The moral wrongness of φ-ing does not by itself have any implications for the normative reasons of the agent. The agent may have decisive

> normative reasons not to φ, sufficient normative reasons, strong but not decisive or sufficient normative reasons, or no normative reasons at all.

An agent has decisive normative reasons not to φ if and only if not φ-ing is all-things-considered normatively required; they have sufficient normative reasons not to φ if and only if not φ-ing is all-things-considered normatively permissible. In Section 3, I argue that MB, together with a further plausible premise linking moral blameworthiness with strong normative reasons, entails the Normativity Thesis. This rules out the No-Entailment Thesis. But MB is consistent with the falsity of both the Overridingness Thesis and the Sufficiency Thesis. I argue that MB does not support the Overridingness Thesis in Section 2 and that it does not support the Sufficiency Thesis in Section 3. There may be other good arguments for the Overridingness Thesis – for example, Kantian arguments (e.g., Korsgaard 1996; Markovits 2014). But there is not a good argument from MB to the Overridingness Thesis.

The argument from MB to the Overridingness Thesis runs as follows. From MB it follows that:

> (P1) If it is morally wrong for an agent to φ, then φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them.

Call such standards 'moral blameworthiness-related standards'. Now add the following claim connecting φ-ing violating moral blameworthiness-related standards and the agent having decisive normative reasons not to φ:

> (P2) If φ-ing violates moral blameworthiness-related standards, then the agent has decisive normative reasons not to φ.

From P1 and P2, it follows that if it is morally wrong for an agent to φ, then that agent has decisive normative reasons not to φ (the Overridingness Thesis).

In the next section, I look at two arguments for P2 and argue that both of them are unpersuasive. To establish P2, it would be enough to establish this simpler claim:

> (S1) If an agent is morally blameworthy for φ-ing, then that agent had decisive normative reasons not to φ.

Now, it is possible to violate moral blameworthiness-related standards without being morally blameworthy, if you have a moral excuse. Hence, at first glance it might seem possible to accept S1 and yet reject P2. Specifically, it might be argued that agents can have sufficient normative reasons for violating moral blameworthiness-related standards when they have a moral excuse. But on further reflection this argument is unsuccessful. Moral excuses show that the agent is not (fully) morally responsible for φ-ing. They do not show that there was anything to be said *in favour of* φ-ing. But then it seems that possessing a moral excuse cannot mark the difference between having or lacking sufficient normative reasons for φ-ing. So, establishing S1 would be enough to establish P2. Both of the arguments for P2 I consider run through S1.

7.2 Two Arguments for P2

The aim of this section is to undermine P2. My aim is not to show that P2 is *false*. Rather, I aim to show that there is not a good argument *from* MB and P2 *to* the Overridingness Thesis. As mentioned earlier, I remain neutral on whether the Overridingness Thesis is correct. Now, if MB and the Overridingness Thesis are true, then P2 is also true. But there is only a good argument from MB and P2 to the Overridingness Thesis if P2 is plausible independently of the Overridingness Thesis. Showing this would require giving an argument for P2 that does not depend on the Overridingness Thesis for its plausibility. In this section, I consider and reject

the two most promising arguments of this kind.[3] I argue that both of these arguments presuppose the Overridingness Thesis at crucial points.

7.2.1 Making Amends

In an early discussion of whether there is a good argument from MB to the Overridingness Thesis, Allan Gibbard claims that there is an apparent incoherence involved in morally blaming someone for φ-ing while being prepared to φ if in exactly like circumstances, and suggests, moreover, that this provides support for the Overridingness Thesis:

> [Are] moral demands always demands of reason? Or can it sometimes make real sense to do things that are morally wrong?... …To judge that it fully makes sense to do a thing is, in effect, not to rule out doing it oneself, if in exactly like circumstances. Now, anger seems incoherent when joined to the thought "If I am in his shoes let me do the same". Likewise with guilt: it seems incoherent when joined to the thought "With the same opportunity, let me do it again". Anger and guilt seem indefensible at the very moment of embracing the act condemned. (1990: 299-300)

(Where Gibbard focusses on anger in general, I focus on resentment and indignation; where he focusses on whether it fully makes sense to do a thing, I focus on whether there are sufficient normative reasons for doing it.) Gibbard's argument has some problematic features. First, it seems to move directly from claims about whether it would be *coherent* to morally blame someone for φ-ing to claims about whether they are morally blame*worthy* for φ-ing. But it is not clear how conclusions about the coherence of moral blame could directly support conclusions about its fittingness. Moreover, Gibbard does not justify or explain the claim that

[3] An argument for P2 I will not consider is Darwall (2006b: 292). For a convincing response to this argument, see Dorsey (2020: 698-700).

the apparent incoherence he points to is a genuine incoherence. Despite these problematic features, Gibbard's argument is worth investigating further. Doing so will allow us to develop a strong argument for S1 that takes inspiration from Gibbard's argument while avoiding its problematic features.

We have seen that guilt motivates agents to make amends; complementarily, resentment and indignation motivate agents to make offenders feel guilty (for the right reasons) and make amends for their conduct. At first glance, the following principle concerning making amends seems plausible:

> *The Repudiation Principle*: Making amends for φ-ing requires *repudiating* φ-ing – that is, committing to not φ-ing again if in exactly like circumstances (and communicating this commitment, in particular to your victim).

There is a kind of practical incoherence involved in being prepared to φ again if in exactly like circumstances while being committed to not φ-ing again if in exactly like circumstances. Insofar as guilt motivates agents to do something that requires them to undertake such a commitment, and resentment and indignation motivate agents to get someone to do something that requires them to undertake such a commitment, combining these emotions with being prepared to φ again if in exactly like circumstances would also seem to yield a kind of tension or incoherence (albeit perhaps an incoherence of a weaker kind). This, I suggest, is the source of the incoherence that Gibbard points to in the passage above.[4]

But there is still a gap between claiming that it would be *incoherent* to feel an emotion in certain circumstances and claiming that it would not be *fitting* to feel it in those circumstances. Fortunately, we will not need to bridge this gap: there is a more direct way of

[4] Gibbard himself characterises guilt chiefly in terms of its connection with making amends (1990: 67-68, 139-140, 146-150).

arguing from the Repudiation Principle to the claim that if an agent is morally blameworthy for φ-ing (i.e., they would be a fitting target of moral blame for φ-ing), then that agent had decisive normative reasons not to φ (S1). To see this, we will need to revisit some points about the fittingness relation made earlier.

In Chapter 5, we saw that an influential account of emotional fittingness holds that it is a matter of *accurate representation*: for an emotion to be fitting is for it to involve an accurate representation of its object (Graham 2014: 392-393; Rosen 2015b: 70-71; Tappolet 2016: 87). I called this view 'Emotional Fittingness as Accurate Representation' (EFAR). I argued against EFAR in Chapter 5, but the arguments of this chapter are intended to run independently. Typically, defenders of EFAR take emotions to have normative representational content. Plausibly, if emotions have such content, then among the factors that are relevant to determining what the content of a given emotion-type is are facts about the kinds of actions they typically motivate (cf. Portmore 2022: 53-54). For example, the claim that fear represents its objects as dangerous is plausible partly because fear typically motivates its subjects to become safe from its objects. Being motivated to become safe from X is an intelligible response to appraising X as dangerous. Now, if the Repudiation Principle is correct, it seems very likely that guilt, resentment, and indignation, insofar as they have normative representational content, include as part of this content a representation of the relevant agents as having had decisive normative reasons not to φ. Otherwise, it is hard to see how being motivated to make amends for φ-ing (or to get someone to feel guilty and make amends for φ-ing), where this involves committing to not φ-ing again if in exactly like circumstances, could be an intelligible response to the appraisals involved in these emotions. The Repudiation Principle therefore supports S1, given an account of fittingness in terms of accurate representation.

Not everyone accepts that fittingness is a matter of representational accuracy – indeed, we saw that there are strong reasons for rejecting this view in Chapter 5. Once EFAR is rejected,

the main alternative views are that fittingness is analysable in terms of reasons (Skorupski 2010; Rowland 2019), or that it is a normatively primitive relation (Chappell 2012; McHugh and Way 2016; Howard 2019). On either of these views, there is a good case for claiming that the Repudiation Principle supports S1. Whether a given emotion is a fitting response to its object, on either of these views, must depend partly on the nature of that emotion – including, surely, its motivational aspect (as I argued in Chapter 5). Now we can ask: what would the fitting objects of guilt have to be like, to merit a response that motivates agents to make amends, where this involves committing to not φ-ing again if in exactly like circumstances? And what would the fitting objects of resentment and indignation have to be like, to merit responses that motivate agents to get offenders to feel guilty and make amends? Plausibly, they would have to be φ-ings such that the agent had decisive normative reasons not to φ. If the agent had only *sufficient* reasons not to φ, or merely a *pro tanto* reason not to φ, then it is hard to see how they could be a fitting target of responses that are tied to *repudiating* φ-ing.

So, it looks as though there is a good argument from the Repudiation Principle to S1, whether fittingness is understood in terms of accurate representation, reasons, or else claimed to be normatively primitive. And, as we saw earlier, establishing S1 would be enough to establish P2. In the remainder of this subsection, I explore the Repudiation Principle further. My aim is not to argue that this principle is false. Rather, whether this principle is plausible depends on whether the Overridingness Thesis is plausible. For there to be a good argument from MB and P2 to the Overridingness Thesis, it would have to be possible to provide an argument for P2 that does not rely on the Overridingness Thesis for its plausibility. The argument for P2 from the Repudiation Principle does not meet this condition, because this principle depends for its plausibility on the Overridingness Thesis.

What is involved in making amends? One element of making amends – arguably, the main element – is providing an apology. Besides apologising, another important element of

making amends is offering compensation, where this is possible and appropriate. There does not seem to be any difficulty in understanding how someone could offer compensation if they do not repudiate φ-ing. For example, Gauguin in *Williams's Gauguin* could provide financial support to his family even if he stands by his decision. So, if the Repudiation Principle is correct, this must be because we cannot provide a full apology for φ-ing if we indicate that we are prepared to φ again if in exactly like circumstances.[5] Many find this an intuitively plausible requirement on full apologies. For example, Nick Smith claims that 'categorical regret is essential to a full apology', where such regret 'entails a promise that the offender will not repeat the offense even under the same conditions and with the same incentives' (2005: 483). If this is right, then the Repudiation Principle must be correct.

But on further reflection, whether full apologies require that the apologiser repudiates φ-ing depends on whether the Overridingness Thesis is correct. To see this, let us return to *Fugitive Son*. Imagine that after a few years the son is discovered and the innocent man who was convicted in his place is released. The innocent man confronts the mother and demands an apology for hiding her son. She says something to the following effect: "I'm sorry for letting you be convicted for my son's crime. I wronged you. If there was any way I could have prevented it from happening without giving him up, I'd have done it. But he asked me to shelter him and he's my son after all". The mother indicates that she stands by the action she performed. If in exactly like circumstances, she would do it again. Yet, consistently with this, she can express the judgments that it was a moral wrong for which she is morally responsible and that it mattered that she wronged him, and she can express her feelings of guilt for having performed it. Is this enough for her apology to be full? Intuitively, this depends on whether she had

[5] What is a 'full' apology? Intuitively, apologies consist of several elements. Uncontroversially, these elements include admission of wrongdoing and acceptance of responsibility. An apology is full when all of the elements that make up apologies are present. It is worth emphasising that a full apology may be qualified in its content. For example, imagine that you tell someone a truth that they need to hear, but tell it in an unnecessarily harsh or tactless manner. Then you may offer a full apology for your tactlessness without apologising at all for telling the relevant truth.

sufficient normative reasons for acting morally wrongly. If she did not, then she has not adequately acknowledged the importance of her moral wrongdoing. But if she did, then she has and her apology is full.

Some may not share this intuition, so it is worth providing further support for the claim that whether full apologies require repudiation depends on whether the Overridingness Thesis is true. Further support is provided by reflection on the kinds of reparative work that apologies can do. As we saw in Chapter 5, one important kind is to withdraw threatening messages that are sent out by unexcused moral wrongdoing, such as the message that moral standards are unimportant and/or that victims of moral wrongdoing and their interests are unimportant (cf. Murphy 1982: 509; Hieronymi 2001: 548-549; Griswold 2007: 55-56). Apologies can withdraw these messages by indicating that their author does not stand by them. A further, related kind of reparative work that apologies can do is to repair damaged trust in others to respect moral standards going forward (Walker 2006: 191-229).

The mother's apology in *Fugitive Son* can do all these kinds of reparative work. It affirms the importance of moral standards and the importance of the victim and their interests (she indicates that, if there were any way she could have prevented the victim from being wrongfully convicted without giving up her son, she would have done this). It could also, in principle, help to repair damaged trust in her to respect moral standards going forward, by showing that she sets great importance by moral standards. Of course, the mother's apology does not demonstrate that she always gives *overriding* importance to respecting moral standards. But again, whether a full apology requires that one demonstrates this must surely depend on the relative importance of morality as compared with other normative domains. If the Overridingness Thesis is false, and the mother has sufficient normative reasons to act morally wrongly, then her apology accurately registers the importance of moral standards and of her victim.

I conclude that the Repudiation Principle depends for its plausibility on the Overridingness Thesis. We therefore cannot appeal to the Repudiation Principle to show that P2 is plausible independently of the Overridingness Thesis.

Before discussing a different argument for P2, let me consider an objection to my criticism of the argument from the Repudiation Principle. The objection is that this argument need not be undermined by showing that the Repudiation Principle depends for its plausibility on the Overridingness Thesis. What the argument could aim to show, it might be suggested, is that there is a cluster of connected claims – MB, P2, the Repudiation Principle, and the Overridingness Thesis – each of which is *prima facie* plausible on its own but which are mutually supportive in such a way that all of them gain additional plausibility. Understood in this way, it is not a problem for the argument if the Repudiation Principle depends for its plausibility on the Overridingness Thesis, since the point of the argument is to demonstrate the interdependence of this cluster of claims.

On this way of construing the argument, its ambitions are relatively modest. It aims to strengthen the confidence of those who already find the Overridingness Thesis plausible, rather than derive this thesis from a set of independently plausible premises. However, there are good reasons for thinking that the argument is unsuccessful in even this modest aim. The discussion of *Fugitive Son* suggests that MB, P2, the Repudiation Principle, and the Overridingness Thesis are not really an interdependent set of claims after all – for it is entirely possible to endorse MB while accepting different understandings of the connection between moral blameworthiness and normative reasons, the requirements of making amends, and the connection between moral wrongness and normative reasons. Moreover, we saw earlier that there is a raft of independent considerations favouring MB, including its intuitive appeal, its ability to explain the common-sense idea that moral wrongs come in different degrees of seriousness, and the fact that it can readily be generalised to account for the unity among different kinds of impermissibility. Hence,

MB, P2, the Repudiation Principle, and the Overridingness Thesis are not really an interdependent set of claims, and we should reject the argument for P2 from the Repudiation Principle even on this re-interpretation of it.

7.2.3 Responsibility and Justice

In this subsection, I consider and reject a different argument for P2. This argument is due to Douglas Portmore (2011: 47-51; 2021: 58-61).[6] Portmore's argument can be paraphrased, with some minor modifications, as follows:

> Q1: An agent is morally blameworthy for φ-ing only if they had the relevant sort of control over φ-ing.
>
> Q2: The agent had the relevant sort of control over φ-ing only if they had the capacity to respond appropriately to the relevant reasons.[7]
>
> (From Q1 and Q2) Q3: An agent is morally blameworthy for φ-ing only if they had the capacity to respond appropriately to the relevant reasons.
>
> Q4: If an agent had sufficient normative reasons to φ, then, by flawlessly exercising their capacity to respond appropriately to the relevant reasons, they could have been led to φ.
>
> Q5: If an agent is morally blameworthy for φ-ing only if they had the capacity to respond appropriately to the relevant reasons, then they cannot be morally blameworthy for φ-ing when, by flawlessly exercising this capacity, they could have been led to φ.

[6] For a different criticism of this argument than the one developed below, see Dorsey (2020: 700-701).
[7] I will sometimes refer to this capacity as 'the capacity for good practical reasoning'.

(From Q3, Q4, and Q5) Q6: An agent is not morally blameworthy for φ-ing if they had sufficient normative reasons to φ.

(From Q6) Q7: An agent is morally blameworthy for φ-ing only if they did not have sufficient normative reasons to φ.

Q8: If an agent did not have sufficient normative reasons to φ, they had decisive normative reasons not to φ.

(From Q7 and Q8) S1: If an agent is morally blameworthy for φ-ing, then that agent had decisive normative reasons not to φ.

As we saw earlier, establishing S1 would be enough to establish P2.

There is much in this argument that is worthy of discussion, but to keep the discussion manageable I will focus on just Q5. In a moment, I will consider some motivations for accepting Q5, but first I will discuss a kind of counter-example to which Q5 may seem to be vulnerable. The kind of case I have in mind is one in which an agent *could have* φ-ed as a result of flawlessly exercising their capacity to respond appropriately to the relevant reasons, but *in fact* φ-ed as a result of bad practical reasoning. Suppose that Linda helps Mark to move into his new apartment. She does this not because she cares about Mark or wants to help him as a friend. She gives no positive weight to these considerations. In fact, she despises Mark and takes the fact that she will benefit him to be a reason not to help him. What motivates Linda is rather her well-supported expectation that if she helps Mark now, he will reciprocate later in some important way. Other things equal, Linda seems morally blameworthy, and this might be thought to show that Q5 is false. However, on further reflection cases of this kind may not be compelling counter-examples to Q5. While Linda is morally blameworthy for *something*, Q5 is only in trouble if she is morally blameworthy for *helping Mark to move into his new apartment*. But a defender of Q5 might deny this. They might claim that Linda is morally

blameworthy for her bad practical reasoning, but not for the action that it leads her to perform (cf. Portmore 2011: 45-46, 50; 2021: 56-57). There is more to be said here, but henceforth I will assume on the strength of this reply that Q5 is not vulnerable to this kind of counter-example.

What, positively, can be said in favour of accepting Q5? Ultimately, Portmore seems to trace the plausibility of this premise, in my view rightly, to the thought that an agent is morally blameworthy for φ-ing only if it would not be unjust to morally blame them for φ-ing (2021: 58-61). He puts the point as follows:

> …it seems that appropriate blame cannot be unjust, and yet it would be unjust to hold an agent responsible on the condition that they have the capacity to be guided by sound practical reasoning and then blame them for acting as they might very well be led to act if they are guided by sound practical reasoning. And since an agent can be led to perform any act that they have sufficient reason to perform when guided by sound practical reasoning, it seems inappropriate to blame them for freely, attributively, and knowledgeably doing what they have sufficient reason to do, as this would be unjust. (2021: 58-59)

In other words, it seems unjust for agents to be morally blamed for acting in ways they could have been led to act by flawlessly exercising a capacity in virtue of which they are eligible for moral blameworthiness in the first place.[8] If an agent is morally blameworthy for φ-ing only if it would not be unjust to morally blame them for φ-ing, it follows that agents cannot be morally

[8] A different interpretation of Portmore's defence of Q5 is that it relies on the claim that moral blameworthiness *itself* cannot be unjust, rather than the claim that an agent is morally blameworthy for φ-ing only if it would not be unjust to morally blame them for φ-ing. In separate work, Portmore argues that if an agent is blameworthy for φ-ing, then they deserve to suffer guilt, remorse, and regret about φ-ing (2022: 63). And he argues further that someone deserves something only if they merit it *as a matter of justice* (2022: 53). This means that blameworthiness in general (including moral blameworthiness) cannot be unjust, because if an agent is blameworthy for φ-ing, then they merit suffering guilt, remorse, and regret about φ-ing as a matter of justice. I argue in fn. 9 that my main criticism of Portmore's argument goes through even on this interpretation.

blameworthy for φ-ing when φ-ing could have resulted from flawlessly exercising such a capacity. For example, suppose that the only way that Tom can spare himself a minor embarrassment is by telling a small lie to Mary. And suppose further that he has sufficient normative reasons either to lie or refrain from lying. Appreciating that he has sufficient reasons to do either of these things, Tom chooses to lie. Would it be just for Mary and others to morally blame Tom for lying? Intuitively, this would be unjust. Tom has exercised his capacity to respond appropriately to the relevant reasons flawlessly, so morally blaming him seems unjust (cf. Portmore 2021: 60-61).

In the remainder of this subsection, I argue that concerns about the justice of moral blame do not in fact support Q5. Given this, Q5 is unsupported and so we do not have good reasons to accept the argument for P2 outlined above. To show that concerns about the justice of moral blame do not support Q5, there are three routes we could take. First, it might be argued that justice simply does not apply to moral blame: like other emotional responses (such as sadness and fear), perhaps moral blame is simply not apt for assessment as just or unjust (cf. Smith 2019). Second, it might be argued that justice is not a constraint on moral blameworthiness: it is not the case that an agent is morally blameworthy for φ-ing only if it would not be unjust to morally blame them for φ-ing (cf. Vargas 2004: 225). For example, imagine that an agent culpably performs a morally wrong action, but later suffers a very great misfortune. Perhaps it could be unjust to morally blame them even though they are morally blameworthy. Third and finally, we might try to meet the concerns about the justice of morally blaming the relevant agents on their own terms. That is, we might concede that moral blame *is* apt for assessment as just or unjust, and that justice *is* a constraint on moral blameworthiness, but argue that it need not be unjust to morally blame people for actions they could have been led to perform by flawlessly exercising their capacity to respond appropriately to the relevant reasons. I will pursue the third approach, but it is worth highlighting that, even if this approach

proves unsuccessful, Q5 may still be unsupported due to one of the other two approaches sketched above.

Note first that, if Q5 is to be part of a non-question begging argument for the Overridingness Thesis, it must be unjust to morally blame someone for φ-ing even under the following circumstances: they could have been led to φ by flawless practical reasoning, but φ-ing was nonetheless morally wrong and they met the conditions of moral responsibility in φ-ing. If it would not be unjust to morally blame someone for φ-ing under these circumstances, then Q5 depends for its plausibility on the Overridingness Thesis (since, if the Overridingness Thesis is true, then the circumstances as described are impossible). Now, there is a good argument from MB and P2 to the Overridingness Thesis only if it is possible to provide an argument for P2 that does not rely on the Overridingness Thesis for its plausibility. If Q5 depends on the Overridingness Thesis for its plausibility, then Portmore's argument fails to meet this condition. I will now argue that under the hypothetical circumstances described above it would be just to morally blame someone for φ-ing. Hence, the argument for P2 outlined above is unsuccessful.

So far, in considering the justice of morally blaming people for actions they could have been led to perform by flawless practical reasoning, I have focussed exclusively on the perspective of potential blamees. Focusing on this perspective reveals a complaint which, if undefeated, seems to sustain a charge of injustice. The complaint is:

> *Complaint*: In performing an action they could have performed as a result of flawless practical reasoning, the potential blamee could have flawlessly exercised one of the capacities in virtue of which they are eligible for moral blameworthiness in the first place. But then it seems unjust for them to be morally blamed for φ-ing.

This complaint provides some support for the charge that it is unjust to morally blame people for actions they could have been led to perform by flawless practical reasoning. However, I will argue that, under the hypothetical circumstances described in the previous paragraph, potential blamers have *justice-relevant interests* in potential blamees being morally blamed, and these interests are strong enough to defeat this complaint and hence defeat the charge of injustice. The interests in question are justice-relevant because they are interests that potential blamers have *as* victims of moral wrongdoing or *as* members of a moral community in which moral wrongdoing has taken place. It bears emphasising that my argument does not presuppose a utilitarian account of moral blameworthiness (on the contrary, I explained in Chapters 1 and 2 that I understand moral blameworthiness in terms of whether agents would be fitting targets of moral blame). It asserts only that the interests of potential blamers are *sometimes* relevant to whether moral blame would be *just*. This is consistent with non-utilitarian accounts of the conditions under which moral blame would be just (such as contractualist accounts). It is also consistent with various views concerning the relation between moral blameworthiness and the justice of moral blame.

The hypothetical circumstances under consideration, recall, are ones in which the relevant agent could have been led to φ by flawless practical reasoning, but φ-ing was nonetheless morally wrong and they met the conditions of moral responsibility in φ-ing. What interests do potential blamers have in agents being morally blamed under these circumstances? Since moral blame comes in both other-directed and self-directed forms, we need to consider the interests they have both in morally blaming this agent themselves and in the potential blamee morally blaming him- or herself. Taking the latter issue first, it is clear that potential blamers have strong interests in the potential blamee morally blaming him- or herself, by feeling guilty. Guilt disposes its subject towards making amends. As we saw in the previous subsection, even agents who acknowledge that they had sufficient normative reasons for

performing an action can be committed to repairing many of the harms it caused, by offering sincere apologies and compensation. Insofar as potential blamers have an interest in moral repair, they have an interest in potential blamees morally blaming themselves. Most obviously, this is an interest that victims have. But it also an interest that third-party blamers have. The repercussions of moral wrongdoing often reach into the wider community, by sending out the message that moral standards are unimportant and undermining trust. Third-party blamers have an interest in potential blamees withdrawing this message and repairing damaged trust by morally blaming themselves and making amends.

Potential blamers also have interests in engaging in moral blame themselves. In particular, they have instrumental interests in doing this. Insofar as feeling and expressing moral blame makes it more likely that potential blamees will morally blame themselves, potential blamers have an interest in engaging in moral blame that derives from their interest in potential blamees morally blaming themselves.

Potential blamers, then, have strong interests in potential blamees being morally blamed under the circumstances we are considering. These circumstances, to recap, are ones in which the potential blamee could have been led to perform the relevant action by flawless practical reasoning, but the action was morally wrong and they met the conditions of moral responsibility in performing it. The fact that the potential blamee could have been led to perform this action by flawless practical reasoning generates a complaint about being subjected to moral blame which, if undefeated, supports a charge of injustice. But given the strong interests that potential blamers have in potential blamees being morally blamed, it seems that this complaint can be adequately met. In particular, it would be unjust for potential blamers to have to bear the cost of the potential blamee's wrongdoing entirely by themselves. Given this, we should conclude that it would be just for potential blamees to be morally blamed under the hypothetical circumstances at issue. If this is right, then concerns about the justice of moral blame do not

support Q5 independently of the Overridingness Thesis, and we do not have good reasons to accept the argument for P2 stated above.

It may help to consider an example. Let me return again to *Fugitive Son*. Suppose, for the sake of argument, that this case meets the conditions specified above: the mother could have been led to shelter her son by flawless practical reasoning, but doing this was nonetheless morally wrong and she met the conditions of moral responsibility in doing it. Given that the mother could have been led to perform this action by flawless practical reasoning, she has a complaint about being the target of moral blame which, if undefeated, supports a charge of injustice. But both the man who was unjustly convicted for her son's crime and third-parties have strong interests in the mother being morally blamed. In particular, this is because they have strong interests in the mother offering moral repair. Given these strong interests, it seems that the mother's complaint can be adequately met. Morally blaming her is just.

Let me now consider an objection. Not all of the interests that potential blamers have in potential blamees being morally blamed are relevant to whether it would be just for them to be morally blamed. For example, there can be cases in which it would be good for potential blamers if potential blamees were *scapegoated* – that is, blamed for something they did not do. But even if these interests are strong enough and the consequences for the potential blamee minor enough that it would be all-things-considered morally right or permissible for them to be morally blamed, this would still be unjust. Scapegoating is a paradigmatic example of unjust blame. Hence it seems that potential blamers have interests in potential blamees being morally blamed that do not bear on whether this would be just. But this casts doubt on my argument that considerations of justice do not support Q5, since this argument appealed to the interests of potential blamers.

However, there are significant disanalogies between scapegoating and the kind of case considered earlier. In this kind of case, the agent really did perform the relevant action. Moreover, the action was morally wrong and they met the conditions of moral responsibility in performing it, even though the practical reasoning that led them to perform it was flawless. The interests I appealed to were interests that potential blamers have *as* victims of moral wrongdoing or *as* members of a moral community in which moral wrongdoing has taken place. These interests do seem relevant to whether it is just for potential blamees to be morally blamed. In particular, they are relevant because it seems unjust for potential blamers to have to bear the costs of the potential blamee's culpable wrongdoing entirely by themselves. But this is what would happen if the potential blamee did not accept responsibility and attempt moral repair. So, the strong interests that potential blamers have in potential blamees being morally blamed (both by themselves and by potential blamers) seem intuitively relevant to the justice of moral blame under these hypothetical circumstances.[9]

I conclude that Q5 in the argument for P2 is not plausible independently of the Overridingness Thesis, and hence that we do not have good reasons for accepting this argument. Neither of the two most promising arguments for P2, then, succeed in showing that P2 is plausible independently of the Overridingness Thesis. Given this, P2 lacks adequate support and hence we should conclude that the argument from MB to the Overridingness Thesis – which runs through P2 – is unsuccessful.

[9] In fn. 8, we saw that there is an alternative interpretation of Portmore's defence of Q5 according to which this defence relies on the claim that moral blameworthiness itself cannot be unjust. On this interpretation, the driving thought behind Q5 is that someone cannot as a matter of justice merit suffering guilt for φ-ing if they could have been led to φ by flawless practical reasoning. The main argument I develop in this section could be re-cast as an argument against the claim that someone cannot as a matter of justice merit suffering guilt for φ-ing if they could have been led to φ by flawless practical reasoning. Again, the crucial move is to argue that, under the hypothetical circumstances at issue, potential blamers have justice-relevant interests in potential blamees suffering guilt and making amends, and these interests are strong enough to defeat concerns about injustice.

7.3 The Normativity Thesis

In 7.1, I contrasted the Overridingness Thesis with the following three theses concerning the connection between moral wrongness and normative reasons:

> *The Sufficiency Thesis*: If it is morally wrong for an agent to φ, then that agent has sufficient normative reasons not to φ.
>
> *The Normativity Thesis*: If it is morally wrong for an agent to φ, then that agent has strong normative reasons not to φ.
>
> *The No-Entailment Thesis*: The moral wrongness of φ-ing does not by itself have any implications for the normative reasons of the agent. The agent may have decisive normative reasons not to φ, sufficient normative reasons, strong but not decisive or sufficient normative reasons, or no normative reasons at all.

In this section, I argue that MB supports the Normativity Thesis but not the Sufficiency Thesis, which is logically stronger.

In 7.2.1, I considered an argument for P2 that turned on reflection on what is involved in making amends. While I argued that this argument is unpersuasive, reflection on what is involved in making amends does, I believe, support the following weaker claim:

> (P3) If φ-ing violates moral blameworthiness-related standards, then the agent has strong normative reasons not to φ.

We saw in 7.1 that MB entails P1:

> (P1) If it is morally wrong for an agent to φ, then φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them.

(That is, if it is morally wrong for an agent to φ, then φ-ing violates moral blameworthiness-related standards). From P1 and P3 it follows that if it is morally wrong for an agent to φ, then that agent has strong normative reasons not to φ (the Normativity Thesis).

P3 is supported by consideration of what is involved in making amends. On reflection, the following principle concerning making amends is highly plausible:

> *The Normative Acknowledgement Principle*: Making amends for φ-ing requires communicating that you had strong normative reasons not to φ, in particular to your victim.

An important element of making amends – arguably, the main element – is providing a full, sincere apology. We came across the claim that sincerely apologising for φ-ing requires communicating to the addressee of the apology that you had strong normative reasons not to φ in Chapters 3 and 6. Now, at first glance it may seem that there are apologies that contravene the Normative Acknowledgement Principle (cf. Chapter 3, fn. 8). Suppose that Jane makes some minor social gaffe. A simple 'I'm sorry' may be all that is required for an adequate apology, and it is not clear that such an apology would convey the message that she had any significant normative reasons not to φ. However, this kind of apology would also not generally be taken to express feelings of guilt (cf. Murphy 2012: 144). Given that our interest is in making amends as an expression of guilt-feelings, we can bracket off these trivial kinds of apologies and focus solely on *serious* apologies. These are apologies that do express guilt-feelings, and which people are typically interested in receiving largely because they take these feelings to stand behind them.

When the apology is a serious one, providing a sincere apology for φ-ing seems impossible if you do not acknowledge having had a significant normative reason not to φ. If you acknowledge breaching moral standards, but profess to regard them as insignificant, like

old-fashioned rules of etiquette, then, intuitively, you do not count as having apologised sincerely. For example, imagine that Gauguin in *Williams's Gauguin* acknowledged being a bad husband or father, but made it abundantly clear that he regarded this as completely unimportant. Intuitively, if he said 'I'm sorry' or 'I apologise' he would not count as apologising sincerely. So, it seems intuitively plausible that the Normative Acknowledgement Principle articulates a requirement on sincere apologies.

Further support for this principle comes from reflection on the reparative effects of apologies. Apologies could not help to repair trust in the apologiser to respect moral standards if the apologiser indicated that they regard these standards as normatively insignificant. Nor could apologies help to withdraw threatening messages sent out by unexcused moral wrongdoing. These include the messages that moral standards are unimportant and that victims of breaches of them are unimportant. Evidently, these threatening messages cannot be withdrawn if the offender professes to regard moral standards as unimportant. So, the Normative Acknowledgement Principle seems well-supported. Significantly, the Normative Acknowledgement Principle does not depend for its plausibility on the Normativity Thesis. Rather, it is supported by independent intuitions about what it takes for an apology to be sincere and reflections on the reparative effects of apologies.

We can get from the Normative Acknowledgement Principle to P3 (if φ-ing violates moral blameworthiness-related standards, then the agent has strong normative reasons not to φ) by means of the following argument:

> T1: Making amends for φ-ing requires communicating that you had strong normative reasons not to φ, in particular to your victim. (The Normative Acknowledgement Principle)

T2: If T1, then, if an agent is morally blameworthy for φ-ing, then they had strong normative reasons not to φ.

S2: If an agent is morally blameworthy for φ-ing, then that agent had strong normative reasons not to φ.

We saw in Section 1 that establishing S1 (if an agent is morally blameworthy for φ-ing, then that agent had decisive normative reasons not to φ) would be enough to establish P2 (if φ-ing violates moral blameworthiness-related standards, then the agent has decisive normative reasons not to φ). By parity of reasoning, establishing S2 would be enough to establish P3 (if φ-ing violates moral blameworthiness-related standards, then the agent has strong normative reasons not to φ).

Support for T2 comes from reflection on the nature of emotional fittingness. We have seen that one influential account of emotional fittingness is EFAR, according to which for an emotion to be fitting is for it to involve an accurate representation of its object. Now, if the Normative Acknowledgement Principle is correct, there are good reasons for thinking that insofar as guilt, resentment, and indignation about an agent's φ-ing have normative content, this content includes a representation of the agent in question as having had strong normative reasons not to φ. As we saw earlier, an important factor in determining the representational content of emotions would be the motivational tendencies associated with them. Guilt, we saw, motivates agents to make amends through such things as apologies and offers of compensation; complementarily, resentment and indignation motivate agents to make offenders hold themselves accountable by feeling guilty (for the right reasons) and making amends for their conduct. If the Normative Acknowledgement Principle is correct, then guilt motivates agents to do something that requires them to acknowledge having had strong normative reasons not to φ. Likewise, resentment and indignation motivate agents to make offenders do something

that requires those offenders to acknowledge having had such reasons. This is strong evidence that these emotions, insofar as they have normative representational content, include a representation of their targets as having had strong normative reasons not to φ. So, on the assumption that emotional fittingness is to be understood in terms of accurate representation, this gives the result that, if the Normative Acknowledgement Principle is correct, then these emotions are fitting only if they are directed towards agents who had strong normative reasons not to φ – and this is just what T2 asserts.

As we saw earlier, if we reject EFAR, the main alternatives are to claim that fittingness is analysable in terms of reasons (Skorupski 2010; Rowland 2019), or else to claim that it is a normatively primitive relation (Chappell 2012; McHugh and Way 2016; Howard 2019). Again, whether a given emotion is a fitting response to its object, on either of these views, must depend partly on the nature of that emotion – including, surely, its motivational aspect. Now we can ask: if the Normative Acknowledgement Principle is correct, what would the fitting objects of guilt have to be like, to merit an emotion that motivates agents to acknowledge having had strong normative reasons not to φ? And what would the fitting objects of resentment and indignation have to be like, to motivate agents to get offenders to acknowledge having had strong normative reasons not to φ? Plausibly, the fitting objects of these emotions would have to be φ-ings such that the agent had strong normative reasons not to φ. Once again, this strongly suggests that if the Normative Acknowledgement Principle is correct, then guilt, resentment, and indignation about an agent's φ-ing are fitting only if the agent had strong normative reasons not to φ – which is what T2 asserts.

So, MB supports the Normativity Thesis. But it might be wondered whether MB also supports the Sufficiency Thesis, which is logically stronger. According to the Sufficiency Thesis, if it is morally wrong for an agent to φ, then that agent has sufficient normative reasons not to φ. This is weaker than the Overridingness Thesis, since it allows for the possibility that

an agent could have sufficient normative reasons to act morally wrongly. But it is stronger than the Normativity Thesis, since it rules out the possibility that an agent could have decisive normative reasons to act morally wrongly.

Whether MB supports the Sufficiency Thesis and not just the Normativity Thesis depends on whether the following claim is true:

> (P4) If φ-ing violates moral blameworthiness-related standards, then the agent has sufficient normative reasons not to φ.

From P1 and P4, it follows that if it is morally wrong for an agent to φ, then that agent has sufficient normative reasons not to φ (the Sufficiency Thesis). I will now argue that we do not have good reasons for accepting P4 in addition to P3. Hence, MB does not support the Sufficiency Thesis in addition to the Normativity Thesis, which is logically weaker.

Reflection on what is involved in making amends does not support P4 in addition to P3. I argued that the following principle concerning making amends is plausible independently of the Normativity Thesis:

> *The Normative Acknowledgement Principle*: Making amends for φ-ing requires communicating that you had strong normative reasons not to φ, in particular to your victim.

Someone might try to argue for P4 by strengthening this principle:

> *The Stronger Normative Acknowledgement Principle*: Making amends for φ-ing requires communicating that you had sufficient normative reasons not to φ, in particular to your victim.

However, as in the case of P2 and the Repudiation Principle, whether the Stronger Normative Acknowledgment Principle is correct depends on whether the Sufficiency Thesis is true. Indeed,

the mother's apology in *Fugitive Son* is most naturally read as expressing the conviction that she had not merely sufficient but decisive normative reasons to shelter her son ("I'm sorry for letting you be convicted for my son's crime. I wronged you. If there was any way I could have prevented it from happening without giving him up, I'd have done it. But he asked me to shelter him and he's my son after all".) As we saw earlier, whether this apology is full and sincere surely depends on the relative importance of morality as compared with other normative domains. Importantly, the mother implicitly acknowledges that she had strong normative reasons to prevent the innocent man from being wrongfully convicted. This allows her apology to have reparative effects, such as repairing damaged trust and withdrawing threatening messages. If she really did have decisive normative reasons to shelter her son, then her apology accurately registers the importance of moral standards and is plausibly full.

So, it seems that we do not have good reasons for accepting P4 in addition to P3. If this is right, then MB does not support the Sufficiency Thesis in addition to the Normativity Thesis. This completes the main argument of this section. I have argued that MB supports the Normativity Thesis, but not the Sufficiency Thesis, which is logically stronger.

7.4 Conclusion

This chapter examined the question of the relation between moral wrongness and normative reasons in light of MB. I argued that MB does not support the Overridingness Thesis, but only the Normativity Thesis, which is logically weaker.

References

Achs, R. & Na'aman. (forthcoming). The subtleties of fit. *Philosophical Studies*.

Anscombe, G. (1958). Modern Moral Philosophy. *Philosophy*, *33*, 1-19.

Archer, A. (2014). Moral Rationalism without Overridingness. *Ratio, 27*(1), 100-114.

Arneson, R. (2015). Basic Equality: Neither Rejectable Nor Acceptable. In U. Steinhoff (Ed.), *Do All Persons Have Equal Moral Worth?* (pp. 30-52). Oxford: Oxford University Press.

Bagley, B. (2017). Properly Proleptic Blame. *Ethics*, *127*(4), 852-882.

Ballard, B. S. (2021). Content and the Fittingness of Emotion. *Philosophical Quarterly*, *71*(4).

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1995). Personal narratives about guilt: Role in action control and interpersonal relationships. *Basic and Applied Social Psychology, 17*, 173-198.

Bealer, G. (1998). A Theory of Concepts and Concept Possession. *Philosophical Issues, 9*, 261-301.

Bell, M. (2013). *Hard Feelings: The Moral Psychology of Contempt.* Oxford: Oxford University Press.

Berker, S. (2022). The Deontic, the Evaluative, and the Fitting. In C. Howard, & R. Rowland (Eds.), *Fittingness.* Oxford: Oxford University Press.

Björnsson, G., & McPherson, T. (2014). Moral Attitudes for Non-Cognitivists: Solving the Specification Problem. *Mind, 123*(489), 1-38.

Brandt, R. (1979). *A Theory of the Good and the Right.* Oxford: Clarendon Press.

Broome, J. (2013). *Rationality Through Reasoning.* Oxford: Wiley-Blackwell.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy, 4*, 73-121.

Burgess, A., Cappelen, H., & Plunkett, D. (Eds.). (2020). *Conceptual Engineering and Conceptual Ethics.* Oxford: Oxford University Press.

Buss, S. (1999). Appearing Respectful: The Moral Significance of Manners. *Ethics, 109*(4), 795-826.

Calhoun, C. (2015). *Moral Aims*. Oxford: Oxford University Press.

Capes, J. A. (2012). Blameworthiness without wrongdoing. *Pacific Philosophical Quarterly, 93*(3), 417-437.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering.* Oxford: Oxford University Press.

Cappelen, H. (2020). Conceptual Engineering: The Master Argument. In H. Cappelen, D. Plunkett, & A. Burgess (Eds.), *Conceptual Engineering and Conceptual Ethics* (pp. 132-151). Oxford: Oxford University Press.

Chappell, R. Y. (2012). Fittingness: The Sole Normative Primitive. *The Philosophical Quarterly, 62*(249), 684-704.

Chappell, R. Y. (2020). Deontic Pluralism and the Right Amount of Good. In D. Portmore (Ed.), *The Oxford Handbook of Consequentialism* (pp. 498-512). Oxford: Oxford University Press.

Clarke, R., & Rawling, P. (forthcoming). True Blame. *Australasian Journal of Philosophy*.

Coates, D. J., & Tognazzini, N. A. (Eds.). (2013). *Blame: Its Nature and Norms.* Oxford: Oxford University Press.

Cohon, R. (1997). The Common Point of View in Hume's Ethics. *Philosophy and Phenomenological Research, 57*(4), 827-850.

Copp, D. (1995). *Morality, Normativity, and Society.* Oxford: Oxford University Press.

Copp, D. (1997a). The Ring of Gyges: Overridingness and the Unity of Reason. *Social Philosophy and Policy, 14*(1), 86-106.

Copp, D. (1997b). Belief, reason, and motivation: Michael Smith's "the moral problem". *Ethics, 108*(1), 33-54.

Copp, D. (2000). Milk, Honey, and the Good Life on Moral Twin Earth. *Synthese, 124*(1), 113-137.

Cuneo, T., & Shafer-Landau, R. (2014). The moral fixed points: new directions for moral nonnaturalism. *Philosophical Studies, 171*, 399–443.

D'Arms, J. (2005). Two Arguments for Sentimentalism. *Philosophical Issues, 15*(1), 1-21.

D'Arms, J. (2022). Fitting Emotions. In C. Howard, & R. Rowland (Eds.), *Fittingness* (pp. 105-129). Oxford: Oxford University Press.

D'Arms, J., & Jacobson, D. (2000a). Sentiment and Value. *Ethics, 110*(4), 722-748.

D'Arms, J., & Jacobson, D. (2000b). The Moralistic Fallacy: On the 'Appropriateness' of Emotions. *Philosophy and Phenomenological Research, 61*(1), 65-90.

D'Arms, J., & Jacobson, D. (2003). The significance of recalcitrant emotion (or, anti-quasijudgmentalism). *Royal Institute of Philosophy Supplements, 52*, 127-145.

D'Arms, J., & Jacobson, D. (2017). Whither Sentimentalism? On Fear, the Fearsome, and the Dangerous. In R. Debes, & K. R. Stueber, *Ethical Sentimentalism: New Perspectives* (pp. 230-249). Cambridge: Cambridge University Press.

D'Arms, J., & Jacobson, D. (2022). The Motivational Theory of Guilt (and Its Implications for Responsibility). In A. B. Carlsson (Ed.), *Self-Blame and Moral Responsibility* (pp. 11-27). Cambridge: Cambridge University Press.

D'Arms, J., & Jacobson, D. (2023). *Rational Sentimentalism.* Oxford: Oxford University Press.

Darwall, S. (1977). Two Kinds of Respect. *Ethics, 88*(1), 36-49.

Darwall, S. (2006a). *The Second-Person Standpoint: Morality, Respect, and Accountability.* Cambridge: Harvard University Press.

Darwall, S. (2006b). Morality and Practical Reason: A Kantian Approach. In D. Copp (Ed.), *The Oxford Handbook of Ethical Theory* (pp. 282-320). Oxford: Oxford University Press.

Darwall, S. (2013a). *Morality, Authority, and Law: Essays in Second-Personal Ethics I.* Oxford: Oxford University Press.

Darwall, S. (2013b). *Honor, History, and Relationship: Essays in Second-Personal Ethics II.* Oxford: Oxford University Press.

Darwall, S. (2017). What Are Moral Reasons? *The Amherst Lecture in Philosophy, 12*, 1-24. Retrieved from http://www.amherstlecture.org/darwall2017/

Darwall, S., Gibbard, A., & Railton, P. (1992). Toward Fin de siecle Ethics: Some Trends. *Philosophical Review, 101*(1), 115-189.

De Hooge, I. E. (2019). Improving Our Understanding of Guilt by Focusing on Its (Inter)personal Consequences. In B. Cokelet, & C. Maley (Eds.), *The Moral Psychology of Guilt* (pp. 131-148). London: Rowman & Littlefield.

De Sousa, R. (2002). Emotional Truth. *Proceedings of the Aristotelian Society, Supplementary Volumes, 76*, 247-263.

Deonna, J. A., & Teroni, F. (2012). *The Emotions: A Philosophical Introduction.* London: Routledge.

Deonna, J. A., & Teroni, F. (2015). Emotions as Attitudes. *Dialectica, 69*(3), 293-311.

Deonna, J. A., & Teroni, F. (2017). Getting Bodily Feelings Into Emotional Experience in the Right Way. *Emotion Review, 9*(1), 55-63.

Deonna, J., & Teroni, F. (2021). Which Attitudes for the Fitting Attitude Analysis of Value? *Theoria, 87*(5), 1099-1122.

Deonna, J., & Teroni, F. (forthcoming). Emotions and Their Correctness Conditions: A Defense of Attitudinalism. *Erkenntnis*.

Deonna, J., Rodogno, R., & Teroni, F. (2011). *In Defense of Shame: The Faces of an Emotion.* Oxford: Oxford University Press.

Dill, B., & Darwall, S. (2014). Moral Psychology as Accountability. In J. D'Arms, & D. Jacobson (Eds.), *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics* (pp. 40-83). Oxford: Oxford University Press.

Dillon, R. (1997). Self-Respect: Moral, Emotional, Political. *Ethics, 107*(2), 226-249.

Dillon, Robin S., "Respect", The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2022/entries/respect/>. (n.d.).

Dorsey, D. (2015). How Not to Argue Against Consequentialism. *Philosophy and Phenomenological Research, 90*(1), 20-48.

Dorsey, D. (2016). *The Limits of Moral Authority.* Oxford: Oxford University Press.

Dorsey, D. (2020). Respecting the Game: Blame and Practice Failure. *Philosophy and Phenomenological Research, 101*(3), 683-703.

Driver, J. (1992). The suberogatory. *Australasian Journal of Philosophy, 70*(3), 286-295.

Driver, J. (2012). Hume's sentimentalist account of moral judgment. In A. Bailey, & D. O'Brien (Eds.), *The Continuum Companion to Hume* (pp. 279-287). London: Continuum.

Driver, J. (2014). Global Utilitarianism. In B. Eggleston, & D. E. Miller (Eds.), *The Cambridge Companion to Utilitarianism* (pp. 166-176). Cambridge: Cambridge University Press.

Fine, K. (1994). Essence and Modality. *Philosophical Perspectives, 8*, 1-16.

Finlay, Stephen and Mark Schroeder, "Reasons for Action: Internal vs. External", The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2017/entries/reasons-internal-external/>. (n.d.).

Fischer, A., & Giner-Sorolla, R. (2016). Contempt: Derogating Others While Keeping Calm. *Emotion Review, 8*(4), 346–357.

Foot, P. (1958). Moral Beliefs. *Proceedings of the Aristotelian Society, 59*, 83-104.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy, 68*(1), 5-20.

Fricker, M. (2016). What's the Point of Blame? A Paradigm Based Explanation. *Noûs, 50*(1), 165-183.

Fricker, M. (2018). Ambivalence About Forgiveness. *Royal Institute of Philosophy Supplement, 84*, 161-185.

Frijda, N. (1986). *The Emotions.* Cambridge: Cambridge University Press.

Frijda, N. (2007). *The Laws of Emotion.* London: Routledge.

Gibbard, A. (1990). *Wise Choices, Apt Feelings.* Cambridge: Harvard University Press.

Gibbard, A. (1992). Moral Concepts: Substance and Sentiment. *Philosophical Perspectives, 6*, 199-221.

Gibbard, A. (2006). Moral feelings and moral concepts. *Oxford Studies in Metaethics, 1*, 195-215.

Giner-Sorolla, R., Kupfer, T., & Sabo, J. (2018). What makes moral disgust special? An integrative functional review. (J. M. Olson, Ed.) *Advances in experimental social psychology*, 223-289.

Goldie, P. (2000). *The Emotions.* Oxford: Oxford University Press.

Graham, P. A. (2014). A Sketch of a Theory of Moral Blameworthiness. *Philosophy and Phenomenological Research, 88*(2), 388-409.

Griswold, C. (2007). *Forgiveness.* Cambridge: Cambridge University Press.

Haidt, J. (2012). *The Righteous Mind*. New York: Penguin.

Harman, G. (2009). Guilt-free morality. *Oxford Studies in Metaethics, 4*, 203-214.

Hart, H. L. (1961/2012). *The Concept of Law*, (3rd ed.). Oxford: Clarendon Press.

Hieronymi, P. (2001). Articulating an Uncompromising Forgiveness. *Philosophy and Phenomenological Research, 62*(3), 529-555.

Hooker, B. (2000). *Ideal Code, Real World.* Oxford: Oxford University Press.

Hooker, B. (2017). What makes a judgement a moral judgement. *Journal of Political Theory and Philosophy, 1*(1), 97-112.

Howard, C. (2018). Fittingness. *Philosophy Compass, 13*(11).

Howard, C. (2019). The Fundamentality of Fit. *Oxford Studies in Metaethics, 14*, 216-236.

Howard, C. (forthcoming). Forever Fitting Feelings. *Philosophy and Phenomenological Research*.

Hughes, Paul M. and Brandon Warmke, "Forgiveness", The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/forgiveness/>. (n.d.).

Hurka, T. (2001). *Virtue, Vice, and Value.* Oxford: Oxford University Press.

Hurka, T. (2019). More Seriously Wrong, More Importantly Right. *Journal of the American Philosophical Association, 5*(1), 41-58.

Isserow, J. (2020). Moral Worth: Having It Both Ways. *Journal of Philosophy, 117*(10), 529-556.

Jackson, F. (1998). *From Metaphysics to Ethics.* Oxford: Oxford University Press.

Jackson, F., & Pettit, P. (1995). Moral Functionalism and Moral Motivation. *The Philosophical Quarterly, 45*(178), 20-40.

Jackson, F., & Pettit, P. (2002). Response-Dependence without Tears. *Philosophical Issues, 12*, 87-117.

Jaworska, Agnieszka and Julie Tannenbaum, "The Grounds of Moral Status", The Stanford Encyclopedia of Philosophy (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), forthcoming URL = <https://plato.stanford.edu/archives/spr2023/entries/. (n.d.).

Jonker, J. (2020). Directed Duties and Moral Repair. *Philosophers' Imprint, 20*(23).

Kamm, F. (2001). Toward the Essence of Nonconsequentialism. In A. Byrne, R. C. Stalnaker, & R. Wedgwood (Eds.), *Fact and Value: Essays on Ethics and Metaphysics for Judith Jarvis Thomson* (pp. 155-181). Cambridge: MIT Press.

Kauppinen, A. (2014). Fittingness and Idealization. *Ethics, 124*(3), 572-588.

Kauppinen, A. (2017). Sentimentalism, Blameworthiness, and Wrongdoing. In K. Stueber, & R. Debes (Eds.), *Ethical Sentimentalism* (pp. 107-132). Cambridge: Cambridge University Press.

Lang, G. (2008). The Right Kind of Solution to the Wrong Kind of Reason Problem. *Utilitas, 20*(4), 472-489.

Laurence, S., & Margolis, E. (1999). Concepts and Cognitive Science. In E. Margolis, & S. Laurence (Eds.), *Concepts: Core Readings* (pp. 3-81). Cambridge: MIT Press.

Laurence, S., & Margolis, E. (2003). Concepts and Conceptual Analysis. *Philosophy and Phenomenological Research, 67*(2), 253-282.

Lewis, D. (1986). *On the Plurality of Worlds.* Oxford: Wiley-Blackwell.

Lewis, D. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society, 63*, 113-137.

Liberto, H. R. (2012). Denying the Suberogatory. *Philosophia, 40*(2), 395-402.

Manela, T. (2015). Obligations of Gratitude and Correlative Rights. *Oxford Studies in Normative Ethics, 5*, 151-170.

Mason, M. (2003). Contempt as a Moral Attitude. *Ethics, 113*(2), 234-272.

Mason, E. (2019). *Ways to be Blameworthy: Rightness, Wrongness, and Responsibility*. Oxford: Oxford University Press.

McDowell, J. (1985). Values and Secondary Qualities. In T. Honderich (Ed.), *Morality and Objectivity* (pp. 110-129). London: Routledge.

McElwee, B. (2017). Supererogation Across Normative Domains. *Australasian Journal of Philosophy, 95*(3), 505-516.

McElwee, B. (2019). The Ambitions of Consequentialism. *Journal of Ethics and Social Philosophy, 17*(2).

McGeer, V. (2019). Scaffolding agency: A proleptic account of the reactive attitudes. *European Journal of Philosophy, 27*(2), 301-323.

McHugh, C., & Way, J. (2016). Fittingness First. *Ethics, 126*(3), 575-606.

McHugh, C., & Way, J. (2022). *Getting Things Right: Fittingness, Value, and Reasons.* Oxford: Oxford University Press.

McKeever, S., & Ridge, M. (2006). *Principled Ethics: Generalism as a Regulative Ideal.* Oxford: Oxford University Press.

McMahan, J. (2009). Intention, permissibility, terrorism, and war. *Philosophical Perspectives, 23*(1), 345-372.

McPherson, T. (2013). Semantic Challenges to Normative Realism. *Philosophy Compass, 8*(2), 126-136.

McPherson, T. (2018). Authoritatively Normative Concepts. *Oxford Studies in Metaethics, 13*, 253-277.

Merli, D. (2002). Return to Moral Twin Earth. *Canadian Journal of Philosophy, 32*(2), 207-240.

Merli, D. (2008). Expressivism and the Limits of Moral Disagreement. *The Journal of Ethics, 12*, 25-55.

Mill, J. S. (1998 [1861]). *Utilitarianism.* (R. Crisp, Ed.) Oxford: Oxford University Press.

Miller, A. (2013). *Contemporary Metaethics: An Introduction, 2nd Edition.* Oxford: Wiley-Blackwell.

Miller, D. E. (2014). "Freedom and Resentment" and Consequentialism: Why 'Strawson's Point' Is Not Strawson's Point. *Journal of Ethics and Social Philosophy, 8*(2), 1-23.

Mitchell, J. (2020). The Irreducibility of Emotional Phenomenology. *Erkenntnis, 85*, 1241-1268.

Moore, A. W. (2006). Maxims and Thick Ethical Concepts. *Ratio, 19*(2), 129-147.

Moore, G. E. (1903). *Principia Ethica.* Cambridge: Cambridge University Press.

Morris, H. (1971). Guilt and Suffering. *Philosophy East and West, 21*(4), 419-434.

Murphy, J. G. (1982). Forgiveness and Resentment. *Midwest Studies in Philosophy, 7*(1), 503-516.

Murphy, J. G. (2012). Remorse, Apology, and Mercy. In J. G. Murphy, *Punishment and the Moral Emotions: Essays in Law, Morality, and Religion* (pp. 129-180). Oxford: Oxford University Press.

Na'aman, O. (2021). The Rationality of Emotional Change: Toward a Process View. *Noûs, 55*(2), 245-269.

Na'aman, O. (2022). What is evaluable for fit? In C. Howard, & R. Rowland (Eds.), *Fittingness.* Oxford: Oxford University Press.

Naar, H. (2021). The fittingness of emotions. *Synthese, 199*(5-6), 13601-13619.

Nussbaum, M. (2001). *Upheavals of Thought.* Cambridge: Cambridge University Press.

Nussbaum, M. (2016). *Anger and Forgiveness: Resentment, Generosity, Justice.* Oxford: Oxford University Press.

Olson, J. (2009). Fitting Attitude Analyses of Value and the Partiality Challenge. *Ethical Theory and Moral Practice, 12*(4), 365-378.

Owens, D. (2012). *Shaping the Normative Landscape.* Oxford: Oxford University Press.

Parfit, D. (2011). *On What Matters, Volume 1.* Oxford: Oxford University Press.

Parkinson, B. (1999). Relations and dissociations between appraisal and emotion ratings ofreasonable and unreasonable anger and guilt. *Cognition and Emotion, 13*, 347-385.

Parkinson, B., & Illingworth, S. (2009). Guilt in Response to Blame from Others. *Cognition and Emotion, 23*, 1589–1614.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life.* Oxford: Oxford University Press.

Pettigrove, G., & Tanaka, K. (2014). Anger and moral judgment. *Australasian Journal of Philosophy, 92*(2), 269-286.

Pettit, P. (1991). Consequentialism. In P. Singer (Ed.), *A Companion to Ethics.* Oxford: Oxford University Press.

Pettit, P., & Smith, M. (2000). Global Consequentialism. In B. Hooker, E. Mason, & D. Miller (Eds.), *Morality, Rules and Consequences: A Critical Reader* (pp. 121-133). Edinburgh: Edinburgh University Press.

Plunkett, D., & Sundell, T. (2013). Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint, 13*(23), 1-37.

Portmore, D. (2011). *Commonsense Consequentialism.* Oxford: Oxford University Press.

Portmore, D. (2021). *Morality and Practical Reasons.* Cambridge: Cambridge University Press.

Portmore, D. (2022). A Comprehensive Account of Blame: Self-Blame, Non-moral Blame, and Blame . In A. Carlsson (Ed.), *Self-Blame and Moral Responsibility* (pp. 48-76). Cambridge: Cambridge University Press.

Prinz, J. (2004). *Gut Reactions.* Oxford: Oxford University Press.

Putnam, H. (1962). It Ain't Necessarily So. *The Journal of Philosophy, 59*(22), 658-671.

Putnam, H. (1975). The Analytic and the Synthetic. In *Mind, Language and Reality: Philosophical Papers* (pp. 33-69). Cambridge: Cambridge University Press.

Queloz, M. (2022). Function-Based Conceptual Engineering and the Authority Problem. *Mind, 131*(524), 1247-1278.

Quine, W. V. (1951). Two Dogmas of Empiricism. *Philosophical Review, 60*(1), 20-43.

Rabinowicz, W., & Rønnow-Rasmussen, T. (2004). The Strike of the Demon: On Fitting Pro-attitudes. *Ethics, 114*(3), 391-423.

Radzik, L. (2009). *Making Amends: Atonement in Morality, Law, and Politics.* Oxford: Oxford University Press.

Roberts, R. (2003). *Emotions.* Cambridge: Cambridge University Press.

Robinson, J. (2017). Emotion as Process. In H. Naar, & F. Teroni (Eds.), *The Ontology of Emotions* (pp. 51-70). Cambridge: Cambridge University Press.

Rosen, G. (2015a). Real Definition. *Analytic Philosophy, 56*(3), 189-209.

Rosen, G. (2015b). The Alethic Conception of Moral Responsibility. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The Nature of Moral Responsibility: New Essays* (pp. 65-88). Oxford: Oxford University Press.

Rossi, M., & Tappolet, C. (2022). Well-Being as Fitting Happiness. In C. Howard, & R. Rowland (Eds.), *Fittingness: Essays in the Philosophy of Normativity.* (pp. 267-289). Oxford: Oxford University Press.

Rossner, M. (2019). Restorative Justice, Anger, and the Transformative Energy of Forgiveness. *The International Journal of Restorative Justice, 2*(3), 368-388.

Rowland, R. (2019). *The Normative and the Evaluative: The Buck-Passing Account of Value.* Oxford: Oxford University Press.

Sayre-McCord, G. (2010). Sentiments and Spectators: Adam Smith's Theory of Moral Judgement. In V. Brown, & S. Fleischacker (Eds.), *The Philosophy of Adam Smith: Essays commemorating the 250th Anniversary of the Theory of Moral Sentiments* (pp. 124-144). (The Adam Smith Review; No. 5). Routledge: London.

Scanlon, T. M. (1998). *What We Owe to Each Other.* Cambridge: Harvard University Press.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame.* Cambridge: Harvard University Press.

Scarantino, A. (2014). The motivational theory of emotions. In J. D'Arms, & D. Jacobson (Eds.), *Moral psychology and human agency: Philosophical essays on the science of ethics* (pp. 156-185). Oxford: Oxford University Press.

Scarantino, Andrea and Ronald de Sousa, "Emotion", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/emotion/>. (n.d.).

Scheffler, S. (1982). *The Rejection of Consequentialism.* Oxford: Clarendon Press.

Scheffler, S. (1985). Agent-Centred Restrictions, Rationality, and the Virtues. *Mind, 94*(375), 409-419.

Schindler, I., Zink, V., Windrich, J., & Menninghaus, W. (2013). Admiration and adoration: their different ways of showing and shaping who we are. *Cognition and Emotion, 27*(1), 85-118.

Schmitt, M., Gollwitzer, M., Förster, N., & Montada, L. (2004). Effects of objective and subjective account components on forgiving. *The Journal of Social Psychology, 144*(5), 465–86.

Schroeder, M. (2007). *Slaves of the Passions.* Oxford: Oxford University Press.

Schroeder, M. (2010). Value and the right kind of reason. *Oxford Studies in Metaethics, 5*, 25-55.

Setiya, K. (2022). What is morality? *Philosophical Studies, 179*, 1113-1133.

Sher, G. (2005). *In Praise of Blame.* Oxford: Oxford University Press.

Shoemaker, D. (2015). *Responsibility from the Margins.* Oxford: Oxford University Press.

Shoemaker, D. (2017). Response-Dependent Responsibility; or, A Funny Thing Happened on the Way to Blame. *The Philosophical Review, 126*(4), 481-527.

Shoemaker, D. (2018). You oughta know: Defending angry blame. In M. Cherry, & O. Flanagan, *The moral psychology of anger* (pp. 67-88). Maryland: Rowman and Littlefield.

Sinclair, N. (2021). *Practical Expressivism.* Oxford: Oxford University Press.

Skorupski, J. (1993). The Definition of Morality. *Royal Institute of Philosophy Supplement, 35*, 121-144.

Skorupski, J. (2010). *The Domain of Reasons.* Oxford: Oxford University Press.

Sliwa, P. (2016). Moral Worth and Moral Knowledge. *Philosophy and Phenomenological Research, 93*(2), 393-41.

Sliwa, P. (2019). The Power of Excuses. *Philosophy & Public Affairs, 47*(1), 37-71.

Smart, J. J. (1961). Free-Will, Praise and Blame. *Mind, 70*(279), 291-306.

Smith, A. (1982). *The Theory of Moral Sentiments.* (D. D. Raphael, & A. L. Macfie, Eds.) Indianapolis: Liberty Classics.

Smith, A. M. (2007). On Being Responsible and Holding Responsible. *The Journal of Ethics, 11*(4), 465-484.

Smith, A. M. (2019). Who's Afraid of a Little Resentment? *Oxford Studies in Agency and Responsibility, 6,* 85-111.

Smith, C. A., Tong, E. M., & Ellsworth, P. C. (2014). The Differentiation of Positive Emotional Experience as Viewed through the Lens of Appraisal Theory. In M. Tugade, M. Shiota, & L. D. Kirby (Eds.), *The Handbook of Positive Emotions* (pp. 11-27). New York: Guildford.

Smith, M. (1994). *The Moral Problem.* Oxford: Wiley.

Smith, M. (2005). Meta-ethics. In F. Jackson, & M. Smith (Eds.), *The Oxford Handbook of Contemporary Philosophy* (pp. 3-30). Oxford: Oxford University Press.

Smith, N. (2005). The Categorical Apology. *Journal of Social Philosophy, 36*(4), 473-496.

Srinivasan, A. (2018). The Aptness of Anger. *The Journal of Political Philosophy, 26*(2), 123-144.

Strawson, P. F. (1962/2003). Freedom and Resentment. In G. Watson (Ed.), *Free Will* (2nd ed., pp. 72-93). Oxford: Oxford University Press.

Stroud, S. (1998). Moral Overridingness and Moral Theory. *Pacific Philosophical Quarterly, 79*(2), 170-189.

Svavarsdóttir, S. (2014). Having Value and Being Worth Valuing. *Journal of Philosophy, 111*(2), 84-109.

Swift, A. (2004). The Morality of School Choice. *Theory and Research in Education, 2*(1), 7-21.

Tadros, V. (2016). *Wrongs and Crimes.* Oxford : Oxford University Press.

Talbert, Matthew, "Moral Responsibility", The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2022/entries/moral-responsibility/>. (n.d.).

Tangney, J. P., & Dearing, R. L. (2002). *Shame and Guilt.* New York: Guilford Press.

Tappolet, C. (2011). Neo-Sentimentalism's Prospects. In C. Bagnoli, *Morality and the Emotions* (pp. 117-134). Oxford: Oxford University Press.

Tappolet, C. (2016). *Emotions, Values, and Agency.* Oxford: Oxford University Press.

Tappolet, C. (2023). *Philosophy of Emotion: A Contemporary Introduction.* London: Routledge.

Thomson, J. J. (1991). Self-Defense. *Philosophy & Public Affairs, 20*(4), 283-310.

Thomson, J. J. (2007). *Normativity.* Chicago: Open Court.

Tierney, H. (2021). Guilty Confessions. *Oxford Studies in Agency and Responsibility, 7*, 182-204.

Tsai, G. (2017). Respect and the Efficacy of Blame. *Oxford Studies in Agency and Responsibility, 4*, 248-275.

Vargas, M. (2004). Responsibility and the Aims of Theory: Strawson and Revisionism. *Pacific Philosophical Quarterly, 85*(2), 218-241.

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility.* Oxford: Oxford University Press.

Walker, M. U. (2006). *Moral Repair.* Cambridge: Cambridge University Press.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments.* Harvard: Harvard University Press.

Wallace, R. J. (2010). Hypocrisy, Moral Address, and the Equal Standing of Persons. *Philosophy and Public Affairs, 38*(4), 307-341.

Wallace, R. J. (2019a). *The Moral Nexus.* Princeton: Princeton University Press.

Wallace, R. J. (2019b). Trust, anger, resentment, forgiveness: On blame and its reasons. *European Journal of Philosophy, 27*(3), 537-551.

Watkins, P. C., Emmons, R. A., Greaves, M. R., & Bell, J. (2018). Joy is a distinct positive emotion: Assessment of joy and relationship to gratitude and well-being. *The Journal of Positive Psychology, 13*(5), 522-539.

Watson, G. (2004). *Agency and Answerability: Selected Essays.* Oxford: Oxford University Press.

Wedgwood, R. (2007). *The Nature of Normativity.* Oxford: Oxford University Press.

Wellman, C. H. (1999). Gratitude as a Virtue. *Pacific Philosophical Quarterly, 80*, 284-300.

Wiggins, D. (1987). A Sensible Subjectivism? In *Needs, Values, Truth: Essays in the Philosophy of Value* (pp. 185–215). Oxford: Blackwell.

Williams, B. (1981). *Moral Luck.* Cambridge: Cambridge University Press.

Williams, B. (1985). *Ethics and the Limits of Philosophy.* London: Fontana.

Williams, B. (1986). Reply to Simon Blackburn. *Philosophical Books, 27*(4), 203-208.

Williams, B. (1993). *Shame and Necessity.* Berkeley: University of California Press.

Williams, B. (1995a). *Making Sense of Humanity.* Cambridge: Cambridge University Press.

Williams, B. (1995c). Replies. In J. E. Altham, & R. Harrison (Eds.), *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams* (pp. 185-224). Cambridge: Cambridge University Press.

Williams, B. (2001). Postscript: Some Further Notes on Internal and External Reasons. In E. Millgram (Ed.), *Varieties of Practical Reasoning* (pp. 91-97). Cambridge: MIT Press.

Wolf, S. (2015). *The Variety of Values.* Oxford: Oxford University Press.

Zimmerman, M. (1997). A Plea for Accuses. *American Philosophical Quarterly, 34*(2), 229-243.

Zimmerman, M. (2008). *Living with Uncertainty: The Moral Significance of Ignorance.* Cambridge: Cambridge University Press.