"I Couldn't Help It!": Essays on Moral Responsibility and Alternative Possibilities

Justin A. Capes
THE FLORIDA STATE UNIVERSITY
COLLEGE OF ARTS AND SCIENCES

“‘I COULDN’T HELP IT!’”
ESSAYS ON MORAL RESPONSIBILITY AND ALTERNATIVE POSSIBILITIES

By
JUSTIN A. CAPES

A Dissertation submitted to the
Department of Philosophy
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Summer Semester, 2011
The members of the committee approve the dissertation of Justin Capes defended on [Type your defense date here].

_________________________________
Alfred Mele
Professor Directing Dissertation

_________________________________
John Kelsay
University Representative

_________________________________
Randolph Clarke
Committee Member

_________________________________
David McNaughton
Committee Member

Approved:

__________________________
J. Piers Rawling, Chair, Department of Philosophy

The Graduate School has verified and approved the above-named committee members.
To “Homer” and “Boogie”, for all your love and support through the years
ACKNOWLEDGEMENTS

I have had some truly wonderful teachers over the years to whom I owe a great intellectual and, in several cases, personal debt, including: Michael Bishop, Randy Clarke, Bill Craig, Linda Damico, Garry Deweese, Doug Geivett, David Hunt, Hugh Hunt, David Horner, David Jones, Rob Koons, Michael McKenna, David McNaughton, Alfred Mele, J.P. Moreland, Piers Rawling, John Roberts, Susan Rouse, Gregg TenElshof, and Gerrit Voogt. I would also like to thank John Kelsay, for his willingness to serve as the university representative on my dissertation committee, and John Fischer, who generously provided valuable feedback on several chapters of this dissertation.

My parents, Earl and Kathy, to whom this dissertation is dedicated, provided both financial and emotional support during my seemingly unending graduate studies. I owe you both a great deal. Thanks also to my brothers Chris and Zac for your friendship and support, and to my girlfriend, Marla, for your endless encouragement and (sometimes unjustified) confidence in me and my philosophical abilities. Finally, I’d like to thank Joe Long, Michael Robinson, and Travis Rodgers from whom I’ve learned much philosophically and whose friendship, cultivated by many days playing disc golf and many nights at “Leon” or “Fermy,” has kept me sane throughout the past four years.
# TABLE OF CONTENTS

ABSTRACT vi

1. THE PRINCIPLE OF ALTERNATE POSSIBILITIES 1
   1.1 Clarifying the Principle 2
   1.2 What’s At Stake? 6
   1.3 The Fallout over Frankfurt 11
   1.4 Coming Attractions 16

2. THE W-DEFENSE 19
   2.1 Frankfurt-style Cases 20
   2.2 Two Arguments 24
   2.3 Blameworthiness and the W-Defense 27
   2.4 Some Unsuccessful Replies 31
   2.5 A Rejoinder to the W-Defense 38
   2.6 A Tentative Observation 42

3. ACTION, RESPONSIBILITY AND ABILITY 44
   3.1 Preliminaries 44
   3.2 Action and Responsibility 47
   3.3 Action and Control 52
   3.4 Action and Inability 54
   3.5 Whose Action? 60
   3.6 Two Implications 62

4. BLAMEWORTHINESS WITHOUT WRONGDOING 64
   4.1 The Objective View 65
   4.2 A Minority Report 71
   4.3 Joining the Minority Report 75
   4.4 The Subjective View 81
   4.5 The Disjunctive View 85

5. ASSESSING THE ALTERNATIVES 86
   5.1 Call It a Draw 86
   5.2 The Traditional Argument for Incompatibilism 90
   5.3 Philosophical Progress 92

REFERENCES 94

BIOGRAPHICAL SKETCH 101
ABSTRACT

According to the Principle of Alternate Possibilities (PAP), a person is blameworthy for what he did only if he could have avoided doing it. This principle figures importantly in disputes about the relationship between determinism, divine foreknowledge, free will and moral responsibility, and has been the subject of considerable controversy for over forty years now. Proponents of the principle have devoted a good deal of energy and ingenuity to defending it against various objections. Surprisingly, however, they have devoted comparatively little effort to developing positive arguments for it, and, with few exceptions, the arguments they have proposed have received little, if any, critical attention. My dissertation is intended to help fill this gap in the literature on PAP.

There are three main arguments for PAP. I critically evaluate each of these arguments, arguing that they are all unsuccessful. Where, then, does that leave PAP? I suggest that, in the absence of any further compelling arguments for or against the principle, debate over it is likely to end in dialectical stalemate. I conclude by highlighting several implications of this suggestion for recent debates about the metaphysics of moral responsibility.
CHAPTER 1

THE PRINCIPLE OF ALTERNATE POSSIBILITIES

The 1964 Lindley Lecture entitled “Human Freedom and the Self” by Roderick Chisholm includes the following passage:

Let us consider some deed, or misdeed, that may be attributed to a responsible agent: one man, say, shot another. If the man was responsible for what he did, then, I would urge, what was to happen at the time of the shooting was something that was entirely up to the man himself. There was a moment at which it was true, both that he could have fired the shot and also that he could have refrained from firing it. And if this is so, then, even though he did fire it, he could have done something else instead. (Chisholm 1969, pp. 24-25)

Here Chisholm clearly articulates the widely accepted thought that a person is morally responsible for what he did only if it was, in some important sense, up to the person himself whether he performed the action in question. Moreover, Chisholm says, if it was up to the person whether he performed the action, then the person could have done something else (or perhaps nothing at all) instead. Together these two claims yield the principle that a person is morally responsible for what he did only if he could have done otherwise. In an influential article, Harry Frankfurt (1969) dubbed this “the principle of alternate possibilities,” and the name stuck.

The principle of alternate possibilities or PAP, for short, has played an important role in discussions of free will and moral responsibility. Like Chisholm, many people have found the principle, or something like it, quite appealing. It apparently explains the exculpatory force of pleas such as “I had no choice” or “I couldn’t help it,” and is often viewed as a plausible way of cashing out the widely held view that free will is required for moral responsibility. On this way of looking at things, if a person couldn’t have avoided doing something, then he did not perform the action in question of his own free
will, in which case he cannot legitimately be held accountable for what he did, however bad or nasty or contrary to accepted standards it might have been. PAP also plays a major role in several philosophical controversies about determinism, divine foreknowledge, free will and moral responsibility. I’ll say more about these controversies shortly.

Whatever plausibility PAP might possess, it is not beyond dispute. There have been numerous efforts in recent years to produce counterexamples to the principle. For what it’s worth, these purported counterexamples seem to me unsuccessful (I’ll briefly explain why later in this chapter and the next), but I will not attempt to defend this claim in any detail here. Instead, I want to focus on a related but slightly different issue. If, as I think, opponents of PAP have failed to refute the principle, so too, I will argue, its defenders have failed to provide compelling reasons to think it is true. The result, I contend, is a classic dialectical stalemate.

My aim in this first chapter is to do a bit of stage setting. I begin by clarifying PAP and explaining why it might matter whether the principle is true. I then give a brief summary of the ongoing debate surrounding it, some of the nuances of which will emerge in subsequent chapters, and I explain the motivation behind this project in light of the current state of that debate. I conclude with a preview of the remaining chapters.

**1.1 Clarifying the Principle**

The first order of business is clarifying PAP. How, exactly, are we to understand the principle? Several issues merit attention in this connection. Let’s start with the topic of responsibility. There is much to say about this topic that is relevant to determining whether principles like PAP are true. However, because my present aim is simply to provide the necessary background for the rest of the dissertation, the discussion that follows will be fairly brief, and will be restricted to those issues most directly relevant to accomplishing this aim. Further issues concerning the nature of moral responsibility, and blameworthiness in particular, will emerge later in chapters two and four.

It is widely agreed that people can be morally responsible for a variety of things other than their actions, including their omissions, the consequences of their actions and omissions, their character traits and various aspects of their mental lives—for example,
their beliefs, desires, intentions, and emotions. PAP, however, is a principle about moral responsibility for actions. I restrict my focus accordingly.

Discussions of moral responsibility tend to focus specifically on blameworthiness. Indeed, some philosophers seem to equate being morally responsible with being blameworthy. On this view, to say that a person is morally responsible for an event or state of affairs just is to say that he is to blame for that event or state of affairs.\(^1\) There is, however, a more inclusive view of responsibility according to which a person can be responsible for something without being blameworthy for it—he might be praiseworthy instead, for example. Which view of responsibility is at issue when it comes to PAP?

Philosophers who equate moral responsibility with blameworthiness will naturally construe PAP as a principle specifically about blameworthiness. However, those who endorse the more inclusive view of responsibility are divided. Most do not restrict PAP to blameworthiness. Others, however, think that there is an important asymmetry between praiseworthiness and blameworthiness in that blameworthiness for actions requires that the agent could have done otherwise, whereas praiseworthiness for actions does not.\(^2\) Finally, there are those who discuss versions of PAP restricted to blameworthiness without addressing versions formulated in terms of praiseworthiness or responsibility more generally.\(^3\) For the most part, this last approach is the one adopted here.\(^4\)

I won’t presuppose any general theory about what it is for a person to be blameworthy for his actions. I will, however, make two important assumptions. First, I assume that if a person is blameworthy for what he did, the action in question reflects ill on the person. It is as if the person has earned a bad mark on his moral record in virtue of what he did. So if a person’s behavior does not reflect ill on him, if it does not occasion a negative mark on his moral record, then he is not blameworthy for what he did. Second, I assume that there is an important link between blameworthiness and certain negative reactive attitudes—resentment, indignation and guilt, in particular. A person is blameworthy if and only if he is the appropriate target of these attitudes.\(^5\)

---

\(^1\) See van Inwagen 1999.  
\(^2\) This “asymmetrical view” is defended by both Nelkin 2008 and Wolf 1990.  
\(^4\) The exception is chapter three. The discussion there needn’t be restricted to blameworthiness.  
\(^5\) Strawson 1962 is often credited with drawing our attention to the connection between responsibility and reactive attitudes. I think he was right about the connection. However, unlike some who take their cue from
I turn now from the topic of morally responsibility to the phrase ‘could have done otherwise’. How are we to understand this phrase as it appears in PAP? According to Alfred Mele, it “is typically given a synchronic reading. That is, PAP is typically understood to assert that a person is morally responsible for what he did at [a time] \( t \) only if, at \( t \), he could have done otherwise then.” As Mele observes, however, when the principle is read that way, it is subject to fairly mundane counterexamples. He cites the case of a drunk driver who, “owing to his being drunk, runs over and kills a pedestrian he does not see.” In an ordinary case of this sort, Mele says, “it is very plausible that…the driver is morally responsible for killing the pedestrian,” even though, “as \( t \) [the time at which the accident happens] draws very near, it is simply too late for the driver to do otherwise at \( t \) than hit and kill the pedestrian” (2006, p. 84).

A lesson some take away from such examples is that ‘could have done otherwise’ in PAP should be read diachronically, not synchronically. That is, PAP should be understood to assert that a person is responsible for what he did at \( t \) only if there was a time \( t^* \), which may or may not be identical to \( t \), at which he could have done otherwise at \( t \). So understood, the principle does not conflict with our judgment that the driver in Mele’s example is responsible for what he did. As Mele observes, in an ordinary case of the sort he has in mind, “the agent was…able at a time significantly earlier than \( t \) to do otherwise at \( t \) than kill the pedestrian then” (p. 85). For example, he could have decided thirty minutes earlier to take a cab and thereby avoided killing the pedestrian at \( t \). The drunk driver case and others like it are therefore not counterexamples to PAP when the principle is read diachronically.

Others draw a slightly different lesson from the likes of Mele’s example. They take such cases to show that PAP should be restricted to direct responsibility. Very roughly, a person is directly responsible for an event or state of affairs provided that he is responsible for it and his responsibility for it does not depend on or derive from his responsibility for some distinct event or state of affairs. If a person’s responsibility for an event or state of affairs is dependent on or derives from his being responsible for something else, then the responsibility for the former event or state of affairs is said to be

---

Strawson’s paper, I am not here committed to the view that all there is to being responsible is being the appropriate object of the relevant attitudes.
indirect. In a normal case of the sort Mele envisions, the driver may indeed be responsible for what he did at \( t \) even though, at \( t \), it was too late for him to do otherwise. However, he is arguably not directly responsible for his actions at \( t \). Had he not been responsible for his decision to drink and drive, then, barring any special assumptions, he wouldn’t have been responsible for hitting and killing the pedestrian at \( t \) either. Evidently, then, his responsibility for hitting and killing the pedestrian at \( t \) depends on his responsibility for deciding to drink and drive in the first place and is thus indirect. The judgment that the driver is indirectly responsible for killing the pedestrian therefore does not conflict with PAP if the principle is restricted to direct responsibility.

One further issue concerning the interpretation of ‘could have done otherwise’ in PAP merits attention. It is perhaps natural to read the phrase this way:

An agent \( S \) could have done otherwise than perform an action \( A = S \) could have done something else (i.e., some other action) instead of \( A \).

On this reading, PAP asserts that a person is morally responsible for what he did only if he could have done something else instead. Thus construed, the principle strikes me as too narrow, however, for it doesn’t cover cases in which the person could have avoided acting as he did, not because he could have performed some alternative action, but rather because he could have simply refrained from performing any action whatsoever. In a case like this, it seems, the person might be responsible for what he did, since he could have avoided doing it, even if he couldn’t have done anything else. For this reason, I focus on variants of PAP formulated in terms of ‘could have avoided’ rather than ‘could have done otherwise’, and I assume that \( S \) could have avoided \( A \)-ing at \( t \) if and only if either (i) \( S \) could have performed an action other than \( A \) at \( t \), or (ii) \( S \) could have simply refrained from performing any action whatsoever at \( t \).

Putting the preceding points together, we can now formulate PAP (or at least the versions of the principle I want to focus on) a bit more precisely as follows:

**PAP:** \( S \) is blameworthy for \( A \)-ing at \( t \) only if there was a time \( t^* \) (which may or may not be identical to \( t \)) at which \( S \) could have avoided \( A \)-ing at \( t \).
In terms of direct responsibility, the principle says

**PAPd**: $S$ is directly blameworthy for $A$-ing at $t$ only if, at $t$, $S$ could have avoided $A$-ing at $t$.

One issue I haven’t yet discussed is the sense of ‘could’ at issue in these principles. I will say more about that in chapter three.

### 1.2 What’s At Stake?

Yesterday, after lunch, I was trying to determine how to spend the rest of my day. Several options occurred to me, but the most appealing were to either run some personal errands or continue working on this dissertation. Following a brief period of deliberation, I decided to postpone running errands so that I could spend the rest of the day writing. Presumably, however, it was within my power to do otherwise; I could have decided instead to get the errands out of the way then and save the writing for another time—or at any rate, so it seems to me.

Many people, I suspect, would not find this little experience of mine particularly remarkable. It certainly strikes me as rather pedestrian. We often find ourselves in situations like this where we must choose between two or more mutually exclusive courses of action. Sometimes, as in the example I just described, the choice is fairly trivial, while at other times it is extremely important. Whether trivial or important, however, we are faced with such choices on a daily basis, and upon reflection it seems to some of us at least that it was within our power to behave differently in the situation than we actually did. In short, it seems to some of us that, in many cases at least, we could have done otherwise. Does anything important hinge on whether those of us who think this are right?

Some philosophers are skeptical about whether it is ever within anyone’s power to do otherwise. Very roughly, *determinism* is the thesis that, given the actual laws of nature, a complete description of the state of the universe at any given time $t$ is consistent
with one and only one possible future subsequent to $t$. Sometimes determinism is taken to include the thesis that every event (except, perhaps, for events occurring at the first moment of time, if there was a first moment) is deterministically caused by prior events, where one event $c$ deterministically causes another event $e$ just in case in every possible world in which the actual laws of nature obtain and in which $c$ occurs, it causes $e$.\(^6\)

**Indeterminism** is the denial of determinism. Here is a classic skeptical argument for the conclusion that no one can do otherwise:

1. Necessarily, if determinism is true, then no one can (is able or has it within his power to) do otherwise.
2. Necessarily, if indeterminism is true, then no one can (...) do otherwise.
3. Either determinism or indeterminism is true.
4. Hence, no one can do otherwise (from 1-3).

Neither premise 1 nor 2 is uncontroversial, but what if they are true? What if this argument, or some other skeptical argument for the same conclusion, is sound?

For many of us, learning that it is never within our power to act otherwise would be surprising and would no doubt occasion a revision of our pre-theoretical views about the nature and extent of the control we possess over our own behavior. But would anything of greater significance follow? In particular, does anything of moral significance depend on whether we can ever do otherwise? If something along the lines of PAP is true, then of course the answer to these questions is ‘yes’. If it is never within anyone’s power to do otherwise and PAP is true, then no one is ever morally responsible for his actions. Thus, if PAP is true, metaphysical debates about whether, and under what conditions, it is within someone’s power to do otherwise have great ethical significance (assuming, as I am, that whether we are sometimes morally responsible for our actions is an important ethical question). If, on the other hand, PAP is false, these metaphysical disputes, which, incidentally, have long been at the center of the free will debate, would lose much, though perhaps not all, of their ethical import.

---

\(^6\) My characterization of determinism follows Clarke 2003, p. 4. For a more detailed account of determinism, see van Inwagen 1983, pp. 2-8; 58-65.
PAP is particularly relevant to the dispute over whether causal determinism is compatible with moral responsibility. It figures importantly in a traditional argument for the incompatibility of determinism and responsibility. In its simplest form, the argument goes like this:

1. Necessarily, if determinism is true, then no one can do otherwise.
2. Necessarily, if no one can do otherwise, then no one is morally responsible for his actions (from PAP).
3. Therefore, necessarily, if determinism is true, no one is ever morally responsible for his actions (from 1 and 2).

More detailed statements of the argument include rather sophisticated arguments in defense of premise 1, but these complications needn’t concern us here.

The traditional response to this argument on the part of compatibilists (those who think determinism and responsibility are compatible) was to challenge premise 1. They argued that a proper understanding of the relevant sense of ‘can’ would reveal that even in a deterministic universe it might be true that people can do otherwise than what they in fact do. It is widely agreed, however, that the project of generating a compatibilist analysis of ‘can’ is a failure. The compatibilist-friendly analyses that were popular in the middle part of the last century all fall victim to fairly decisive counterexamples and recent attempts to improve on these analyses evidently fail as well. For this reason, many compatibilists now prefer to challenge premise 2 and the principle—viz., PAP—on which it is based.

The question of whether PAP is true is also relevant to a classic theological puzzle about whether infallible divine foreknowledge is compatible with human freedom and responsibility. Let us refer to this puzzle as the problem of theological fatalism. The theological fatalist maintains that, necessarily, if God exists and has infallible foreknowledge of future events, including future human actions, then none of us ever acts of our own free will. And if, as is commonly assumed, moral responsibility requires free

---

7 Versions of this argument are defended by Chisholm 1964, Lamb 1977 and van Inwagen 1983, among others.
8 For a helpful discussion of these issues, see Clarke 2009.
will, it follows from the theological fatalist’s position that, necessarily, if God exists and has infallible foreknowledge, then no one is ever morally responsible for his actions.

What is it to say that someone has free will? According to Peter van Inwagen, it is to say that “very often, if not always, when [the person] has to choose between two or more mutually incompatible courses of action…each of these courses of action is such that he can, or is able to, or has it within his power to carry it out (1983, p. 8). On van Inwagen’s view, then, having free will implies that it was within the agent’s power to do something other than what he actually did. This view of free will is common, though by no means uncontroversial. Here are two further representative examples of it:

If a person \( S \) is free with respect to a given action, then he is free to perform that action and free to refrain…It is within his power, at the time in question, to perform the action, and within his power to refrain. (Plantinga 1974a, pp. 165-166)

\([S] \) is free at \( T \) with respect to performing \( A \) =df It is in \([S] \)’s power at \( T \) to perform \( A \), and it is in \([S] \)’s power at \( T \) to refrain from performing \( A \). (Hasker 1989, p. 66)

This view of free will is among the assumptions that generate the problem of theological fatalism.

Why does the theological fatalist think that infallible foreknowledge is inconsistent with human freedom and responsibility? In its simplest form, the argument looks like this:

1. Necessarily, if God exists and has exhaustive and infallible foreknowledge and \( S \) does \( A \) at \( t \), then it was not within \( S \)’s power to avoid \( A \)-ing at \( t \).
2. Necessarily, if it was not within \( S \)’s power to avoid \( A \)-ing at \( t \), then \( S \) did not \( A \) at \( t \) of his own free will (from the definition of free will).
3. Necessarily, if \( S \) did not \( A \) at \( t \) of his own free will, then \( S \) is not morally responsible for \( A \)-ing at \( t \).
4. Therefore, necessarily, if God exists and has exhaustive and infallible foreknowledge and $S$ does $A$ at $t$, then $S$ did not $A$ at $t$ of his own free will and is not morally responsible for $A$-ing at $t$ (from 1-3).  

As was the case with the traditional argument for the incompatibility of determinism and responsibility presented above, more detailed statements of the present argument include sophisticated defenses of premise 1, but again, the present formulation of the argument will suffice for present purposes.

Standard attempts to solve the problem of theological fatalism involve challenging the defense of premise 1. Another response on the part of some theists concedes that the argument is sound and thus jettisons either divine foreknowledge or human freedom and responsibility. Most theologians have been loath to give up on freedom and responsibility. Some, however, have been willing to abandon exhaustive divine foreknowledge, arguing that the traditional conception of God, which includes the claim that God is essentially omniscient, does not, in fact, require it. According to this view (sometimes known as ‘open theism’), divine omniscience involves knowing everything that can be known. Hence, since God, according to the theological fatalist, cannot foreknow future free actions, his lack of foreknowledge about what people will freely do in the future does not count against his being omniscient. A third sort of response involves denying PAP.

Suppose for the moment that PAP is false. In that case, a person could be morally responsible for what he did even if he couldn’t have avoided doing it and thus, if premise 2 of the argument for theological fatalism is true, even if he did not have free will with respect to the action in question. So if PAP is false, one could solve the problem of theological fatalism by accepting premise 2 (the claim that free will requires the ability to do otherwise) and rejecting premise 3 (the claim that free will is required for responsibility). Alternatively, those who reject PAP might accept premise 3 and argue that, since PAP is false, the conception of free will advocated by van Inwagen, Plantinga, Hasker and others should be rejected. Contrary to what these authors claim, an opponent

---

9 For more detailed formulations of the argument, see Pike 1965 and Hasker 1989.
10 For this latter sort of response, see Swinburne 1977 and van Inwagen 2006.
of PAP might insist, a person can be morally responsible for what he did, and thus, given premise 3, can act of his own free will, even if he couldn’t have avoided doing what he did.

Those who favor this second strategy might replace the view of freedom that generates the problem of theological fatalism with what is sometimes referred to as an ‘actual sequence’ account of free action. According to actual sequence theories, what is essential to free action is not that the agent could have done otherwise than he actually did but, rather, how his action was actually produced. On this way of thinking of things, only factors that influence how the action in question is produced are relevant to whether it is a free action. If the action was the result of undue force or external coercion, for example, then the actual sequence theorist might not count it as free. Now, even if infallible divine foreknowledge is incompatible with the ability to do otherwise, because God’s foreknowledge does not cause or otherwise compel our actions, there is no reason to think that it is incompatible with our freedom and responsibility, given an actual sequence account of these things.\textsuperscript{11}

1.3 The Fallout Over Frankfurt

My primary aim in the preceding section was to explain why the debate over PAP is philosophically significant. I now turn now to the debate itself. It was sparked by an influential article of Harry Frankfurt’s (1969).\textsuperscript{12} In that paper Frankfurt produced a series of examples that culminate in what he and many others think is a counterexample to the principle. Much of the subsequent literature on these issues has consisted in various attempts to repair, refine or respond to Frankfurt’s purported counterexample. I will discuss Frankfurt’s criticism of PAP in a bit more detail in chapter two, but for now a brief sketch of his argument and the controversy that ensued will suffice.

Consider the following example:

\textsuperscript{11} Hunt 1996, 1999 and 2003 explores this sort of solution to the problem of theological fatalism.

\textsuperscript{12} The roots of Frankfurt’s argument can be found in Augustine’s treatment of the relationship between human freedom and divine foreknowledge in On the Free Choice of the Will and in Locke’s discussion of free will in the section ‘Of Power’ in An Essay Concerning Human Understanding. See Hunt 1996 and 1999 for a helpful discussion of Augustine’s position. For discussion of Locke’s views, see Hoffman 2005. For a different argument against PAP, see Graham 2008.
MORNING BREW: A maniacal villain has covertly slipped a powerful, mind altering drug into Al’s morning coffee that produces in him an irresistible urge to strangle his friend, Bob. Powerless to resist the urge, Al rushes over to Bob’s house and strangles him.

It seems clear that Al is not morally responsible for strangling Bob. We cannot legitimately blame him for doing so, nor is he deserving of punishment on the basis of what he did. But why not? Why does Al get off the hook? An initially plausible answer to this question is that he couldn’t help doing what he did. Through no fault of his own, he found himself in circumstances in which he couldn’t avoid strangling Bob, and thus, according to PAP, he is not responsible for doing so.

Frankfurt insists that PAP may seem initially appealing because, in seeking illustrations of the principle, we naturally tend to think only of cases like MORNING BREW in which the factors that bring it about that a person performs the action in question are the very same factors that prevent the person from doing otherwise. Other such cases might include those in which the agent is forced against his will, is the victim of covert manipulation or suffers from some psychological disorder. In cases like these the agent typically is not morally responsible for his actions. No mystery, then, why those who focus solely on such cases might be attracted to PAP. However, Frankfurt claims that there can be situations—call them ‘FR-situations’—in which the factors that make it so that a person cannot do otherwise are distinct from those that bring it about that the person acts as he does, and in situations of this sort, Frankfurt thinks, the person may well be responsible for what he did even though he couldn’t have avoided doing it.

To illustrate an FR-situation, Frankfurt appeals to the following example:

Black…wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something other than what he wants
him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones’s initial preferences and inclinations, then, Black will have his way…. [However] Black never has to show his hand because Jones, for reasons of his own decides to perform and does perform the very action Black wants him to perform. (1969, pp. 835-836)

To fill out the example, let’s suppose that what Black wants Jones to do, and what Jones actually does, is to shoot and kill a third man named Smith, and let us refer to the case as ‘SHOOT SMITH’.

Frankfurt believes that SHOOT SMITH is an FR-situation, and he argues that such cases are, or at least provide the basis for, counterexamples to PAP. Evidently, Jones couldn’t have done otherwise than kill Smith. His only options, it seems, were to shoot and kill Smith on his own or as a result Black’s intervention. Either way, though, he does exactly what Black wants him to do. However, because Jones acted on his own, without any help from Black, there is pressure to say that he is, or at least could be, morally responsible for shooting and killing Smith nonetheless.

Frankfurt’s argument sparked a heated debate, one that has yet to approach anything resembling a resolution. Defenders of PAP have of course denied that his example or those modeled after it—the “Frankfurt-style cases,” as they have come to be known—are genuine counterexamples to PAP. Typically, their reasons for denying this come in one of three forms. Some have argued that Frankfurt-style cases are conceptually incoherent. Others insist that employing them in an argument against PAP is dialectically inappropriate, for in doing so one inevitably winds up begging the question against incompatibilists. Still others have argued that, although the agent in Frankfurt-style cases is perhaps morally responsible, he in fact had access to relevant alternative possibilities. This is sometimes referred to as the ‘flicker of freedom’ defense.

---

13 E.g., see Alvarez 2009.
15 The term ‘flicker of freedom’ was coined by Fischer 1994 to refer to residual alternative possibilities that remain in Frankfurt-style cases.
There are several different variations of this third strategy, but to my mind the most promising versions appeal to Frankfurt’s observation that, although it is not up to Jones which action he performs, “it is in a way up to him whether he acts on his own or as a result of Black’s intervention” (1969, p. 836). Several defenders of PAP have seized on this feature of Frankfurt’s example in an effort to salvage the principle, or something close to it.\(^{16}\) Their claim is that what Jones is responsible for, or at least what he is most directly responsible for, is not shooting Smith per se but rather for shooting-Smith-on-his-own. Since Jones could have avoided shooting-Smith-on-his-own—a point even Frankfurt is apparently willing to concede—we do not have a counterexample to PAP after all.\(^{17}\)

What are we to make of these responses to Frankfurt’s argument? The first response, according to which Frankfurt-style cases are incoherent, is the least promising of the three. I consider and reject a version of this strategy in chapter three. The second response, according to which employing Frankfurt-style cases in an argument against PAP inevitably begs the question against incompatibilists, is fairly weak as well. At best it shows that Frankfurt’s argument won’t convince those already committed to incompatibilism, but it doesn’t give the compatibilist or a reasonable agnostic about these issues any reason to reject the argument.\(^{18}\) The third response, according to which the agent in the examples is responsible for *something* he did but could have avoided doing, strikes me as the most promising. If successful, it shows that we can accommodate our intuition that the agent in Frankfurt-style cases is morally responsible without jettisoning the likes of PAP.

Why might someone think that Jones is responsible for shooting-Smith-on-his-own but not for shooting Smith *simpliciter*? The argument, I think, goes something like this. Prior to anything Jones does on his own accord, circumstances over which he has no control are in place which guarantee that, whatever else he does, if he does anything at

---


\(^{17}\) Given the role PAP plays in the standard argument for incompatibilism mentioned earlier, it is perhaps unsurprising that many of those keen on defending the principle are incompatibilists. But it’s important to note that some compatibilists still accept versions of PAP, while some incompatibilists reject it. Compatibilists who accept PAP include Fara 2008 and Vihvelin 2004. Incompatibilists who reject PAP include Pereboom 2001, Hunt 2000 and 2005, Stump 1990, 1996 and 2003, and Zagzebski 1991.

\(^{18}\) A similar point is developed in detail by Haji and McKenna 2004.
the time in question, he will certainly shoot Smith. In this respect, cases like SHOOT SMITH are similar to cases like MORNING BREW. Thus, it will be argued, Jones should not be blamed for shooting Smith any more than Al should be blamed for strangling Bob. To be clear, the point isn’t just that Jones is in circumstances that ensure he will shoot Smith but, rather, that his being in these circumstances is something over which he has absolutely no control. Together these facts are enough, some defenders of PAP will argue, to establish that Jones is not blameworthy for shooting Smith. However, since it was up to Jones whether he shot Smith on his own or as a result of Black’s intervention, we can blame him for shooting-Smith-on-his-own.¹⁹

Whatever plausibility this defense of PAP may possess (and I think it has a good deal), it can only go so far. It can show how we might accommodate our intuitions about Frankfurt-style cases without jettisoning the likes of PAP, but it does not give us any reason to accept the principle. So can defenders of the principle go further? Can they show that there are good reasons for thinking that PAP is true? Given the energy and ingenuity they have devoted to defending it against Frankfurt-style cases in recent years, one might expect that they would have devoted a similar amount of effort to providing positive arguments for it. But as it turns out, this is not the case. Proponents of PAP have devoted surprisingly little effort to developing positive arguments for the principle, and, with a few exceptions, the arguments they have suggested have received little critical attention. A central aim of this dissertation is to fill this gap in the literature on PAP.

The main question I am interested in, then, is whether there are any compelling reasons to think PAP is true. This is an important question. To see why, consider an analogy with a rather different philosophical debate—the debate between theists and non-theists over the existence of God. The problem of evil is the central argument against theism. Suppose the theist can show that this is not a good argument against the existence of God. That would be an interesting and important conclusion, but it certainly wouldn’t settle the dispute in favor of theism. What would settle matters in favor of theism? A compelling argument for the existence of God might turn the trick. That in and of itself is

¹⁹ A slightly different approach is to concede that Jones is responsible for shooting Smith even though he couldn’t have done otherwise, but to insist that his responsibility for this action derives from his responsibility for shooting Smith on his own and is thus indirect. Robinson (forthcoming) suggests this strategy. Those who favor this approach might be willing to concede that some Frankfurt-style cases are counterexamples to PAP but not to PAPd.
a reason for those interested in the question of whether God exists to carefully consider theistic arguments. Now, it may be, as even some theists are willing to concede, that none of the traditional arguments for God’s existence are compelling. Even so, it would certainly be strange, and they would certainly be remiss, if philosophers interested in the issue neglected to consider those arguments. Moreover, even if all arguments for God’s existence are indeed failures, we may nevertheless be able to learn much philosophically from studying them. As Alvin Plantinga observes, part of the explanation for the enduring interest in the ontological argument for God’s existence has to do with the fact that “many of the most knotty and difficult problems in philosophy meet in this argument” (1974b, p. 85). There is, then, value in studying the theistic arguments even if they don’t provide us with compelling reasons to believe in God.

Much the same remarks apply, *mutatis mutandis*, to the debate over PAP. Suppose that defenders of the principle can show that Frankfurt-style cases do not give us compelling reason to reject it. That would be an interesting and important conclusion, but it certainly wouldn’t settle the dispute in favor of PAP. What would settle matters in favor of PAP? A compelling argument for the principle might turn the trick. That in and of itself is a reason to carefully consider arguments for PAP. Now, perhaps, as I will argue, none of the main arguments for the principle are successful. Even so, it would certainly be strange, and we would certainly be remiss, if those of us interested in the question of whether PAP is true neglected to consider these arguments. Moreover, even if they are all unsuccessful, we may nevertheless gain important philosophical insights from studying them. As we shall see, these arguments raise issues about the nature of moral responsibility, obligation and human action that are important and interesting in their own right. There is, then, value in studying arguments for PAP even if they don’t provide us with compelling reasons to accept the principle.

### 1.4 Coming Attractions

That’s all I have to say by way of introduction to the current project. The remainder of the dissertation plays out as follows. In chapters two, three and four, I evaluate the three
main arguments for PAP recently put forward and defended in the literature. I contend that each of these arguments is unsuccessful. In the fifth and final chapter, I consider the implications of this conclusion for the debate over PAP and for related controversies concerning the metaphysics of free will and moral responsibility.

In chapter two, I examine an argument for PAP that hinges on a theory of blameworthiness according to which a person is blameworthy for what he did if and only if it would have been reasonable to expect or demand that he not to do it. On the assumption that it is reasonable to expect someone not to do something only if the person can avoid doing it, the theory entails PAP. Unfortunately, the theory of blameworthiness on which this argument turns has little to recommend it.

A second argument for PAP is based on the novel idea that performing an action requires the ability to refrain from performing it. If this idea is right, then a fortiori performing a free action for which one can be morally responsible requires that one could refrain from performing it. This argument is the subject of chapter three. Contrary to what proponents of this argument insist, performing an action does not require that one could have done otherwise. Agents who suffer from various psychological disorders and addictions may engage in compulsive behaviors that, given their condition, they cannot avoid. This behavior may nevertheless be among their intentional actions. Moreover, even normal agents might occasionally perform actions from which they are unable to refrain, if, for example, they lack and cannot acquire motivation to do otherwise. This is true even if it turns out that it is very often within our power to do otherwise.

Perhaps the central argument for PAP turns on the venerable deontic maxim that ‘ought’ implies ‘can’. Given a pair of auxiliary assumptions, PAP is a corollary of this maxim. One response to this argument is to reject the ‘ought’ implies ‘can’ principle. Another is to reject one (or both) of the auxiliary assumptions. I advocate the latter response. One advantage of this strategy is that it allows one to remain neutral about whether ‘ought’ implies ‘can’.

---

20 Zimmerman 2003 considers several other possible arguments for PAP, but as far as I am aware, only those discussed here have been worked out in any detail.
21 Fischer 2006 takes this first route.
22 Zimmerman 1997 takes this second route.
In chapter four I consider the relationship between blameworthiness and wrongdoing. Among the theses I defend in that chapter is that a person can be blameworthy for what he did even if he didn’t do anything objectively wrong. If this is right, then the central argument for PAP is unsuccessful, for one of the auxiliary assumptions needed to derive PAP from the ‘ought’ implies ‘can’ maxim is the claim that blameworthiness requires objective wrongdoing.

If the main arguments for PAP are unsuccessful, where does that leave PAP? My aim in chapter five is to address this question. I suggest that, in the absence of any further arguments for or against PAP, debate over the principle ends in dialectical stalemate. I conclude by highlighting several implications of this suggestion for metaphysical disputes about the relationship between determinism, free will and moral responsibility.
CHAPTER 2

THE W-DEFENSE

Recall that, according to Frankfurt (1969), there may be circumstances that make it impossible for a person to avoid performing a certain action that play no role in bringing it about that the person performs that action. Frankfurt contends that in circumstances of this sort the person may be morally responsible for what he did, despite the fact that he couldn’t have avoided doing it. If circumstances of this sort are indeed possible, and if Frankfurt is right that a person who finds himself in these circumstances might be responsible for what he does even though he cannot avoid doing it, then PAP is false.

As I mentioned in chapter one, there has been a great deal of critical discussion of Frankfurt’s argument in recent years. Most of the debate has focused on the question of whether the sorts of circumstances envisioned by Frankfurt (FR-situations) are possible. Until quite recently, however, very little attention has been devoted to the question of whether an agent in FR-situations could indeed be morally responsible for what he has done, as Frankfurt claims. One critic of Frankfurt’s argument who has addressed this question is David Widerker (2000, 2003, and 2005). Widerker contends that even if FR-situations are possible, Frankfurt’s argument does not succeed in refuting versions of PAP restricted to blameworthiness, since agents in FR-situations could never be blameworthy for their actions. For reasons that will become clear shortly, he refers to his argument as the what-should-he-have-done-defense or the W-defense, for short.

The W-defense is important for at least two reasons. First, as I explain in a bit more detail later, the dispute over the possibility of FR-situations has proven rather difficult to resolve, and if current trends are any indication, no resolution of the matter is forthcoming. However, the W-defense suggests a way of advancing the debate over PAP that would allow one to bypass the thorny metaphysical issues involved in assessing the possibility of FR-situations. Second, the W-defense raises interesting questions regarding the nature of moral blame, and as I will argue, one’s assessment of both Frankfurt’s
argument against PAP and the W-defense will depend to some extent on what one thinks is involved in blaming someone.

In this chapter, I examine the W-defense in some detail, along with three recent replies to it by defenders of Frankfurt’s argument, including Frankfurt himself. I contend that each of these replies plays into the hands of the PAP-defender in one way or another. However, I do not come to champion the W-defense. Instead, I develop my own criticism of it by calling into question an assumption regarding the nature of moral blame that is at the very heart of Widerker’s argument. I begin by briefly locating the W-defense within the larger dialectic surrounding Frankfurt’s argument.

2.1 Frankfurt-style Cases

Frankfurt’s argument against PAP relies on the following two assumptions:

**FR**: There can be circumstances that make it impossible for an agent \( S \) to avoid performing a certain action \( A \) that play no role in bringing it about that \( S A \)-s.

**M**: In circumstances of this sort (i.e., FR-situations) \( S \) might very well be blameworthy for \( A \)-ing even though he couldn’t have avoided \( A \)-ing.

It follows from these assumptions that PAP is false.

As I mentioned earlier, the controversy surrounding Frankfurt’s argument has focused almost entirely on the truth of FR and whether the circumstances described therein are indeed possible. This is where the Frankfurt-style cases become relevant. To illustrate an FR-situation, Frankfurt offered this example, which I introduced in chapter one.

**SHOOT SMITH**: Black…wants Jones to [shoot Smith]. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such
things) that Jones is going to decide to do something other than [shoot Smith]. If it
does become clear that Jones is going to decide to do something else, Black takes
effective steps to ensure that Jones decides to do, and that he does do, what he
wants him to do. Whatever Jones’s initial preferences and inclinations, then,
Black will have his way…. [However] Black never has to show his hand because
Jones, for reasons of his own decides to [shoot Smith and does [shoot Smith].
(1969, pp. 835-836)

This is supposed to be a FR-situation: given Black’s presence, intentions and abilities,
Jones apparently cannot avoid shooting Smith, for if he does not shoot Smith on his own
accord, Black will intervene to make sure that he shoots Smith nevertheless. However, in
the actual sequence of events, Black does not have to intervene. He does not force, coerce
or otherwise induce Jones to act because Jones shoots Smith entirely of his own accord. It
seems, then, that, although Black’s presence, abilities and intentions make it so that Jones
cannot avoid shooting Smith, these factors do not play a role in causing or otherwise
bringing it about that Jones shoots Smith.

Many PAP-defenders have challenged Frankfurt’s argument by questioning
whether this example really is an FR-scenario. If it is not, then Frankfurt will have failed
to establish a key premise of his argument. And there are indeed a number of problems
with the example as it stands. For starters, we might wonder how Black could possibly
determine what Jones is going to do before he actually does it. A common answer to this
question involves elaborating on the example as follows. Imagine that Black is nefarious
neurosurgeon who has developed sophisticated bran-scan device that allows him to read
Jones’s thoughts and control his behavior. Suppose this machine enables Black to
observe Jones’s neural states from afar and that on the basis of his observations he is able
to detect a certain neural pattern or sign that he knows is somehow correlated with
decisions to perform the sort of action he wants Jones to perform. Thus, if at any point
during the deliberative process Jones exhibits this sign, Black will know that intervention
on his part is unnecessary. This way of embellishing the case is somewhat fanciful, to be
sure, but all Frankfurt needs in order make his point is a metaphysically possible
scenario, and this version of the example, fanciful as it may be, might seem to turn the trick. But there is a problem.

Several defenders of PAP have argued that augmenting Frankfurt’s example in this way may solve the original problem about how Black can know in advance what Jones is going to do, but only at the expense of raising another, more complex problem for Frankfurt’s overall argument. Note that the relation between the prior sign (or that of which it is indicative) and Jones’s action is either deterministic or it is not. If it is, then claiming that Jones is blameworthy for what he did begs the question against incompatibilists about determinism and moral responsibility. If, on the other hand, the relation between the prior sign (or that of which it is indicative) and Jones’s action is not deterministic—if, for example, the sign is merely a reliable, but ultimately fallible, indicator of what Jones is going to do—then the presence of the sign does not guarantee that Jones is going to shoot Smith, in which case Black cannot ensure that Jones will shoot Smith except by intervening and causing Jones to shoot Smith. To see this, suppose the sign Black actually used was a reliable, but ultimately fallible, indicator of what Jones would do. Since he saw the sign, Black did not intervene. But given that the sign is reliable but still fallible, its presence did not guarantee that Jones would shoot Smith. So, since Black did not intervene, and barring any further assumptions that might suggest that Jones couldn’t have avoided shooting Smith, what reason is there to think that Jones couldn’t have avoided doing what he did?

It seems, then, that Frankfurt’s argument now faces a dilemma. If the relation between the prior sign (or something of which it is indicative) and Jones’s action is deterministic, then Frankfurt cannot claim that Jones is blameworthy for shooting Smith without begging the question against incompatibilists, nor can he claim that the example is an FR-situation, for if there is a deterministic relation between prior events and Jones’s action, then the incompatibilist can argue that there are features of the circumstances that play a role in producing Jones action that are also among the circumstances that make it so that he cannot do otherwise. But if the relation between the prior sign (or that of which it is indicative) and Jones’s action is not deterministic, then we no longer have any reason to suppose that the circumstances in which Jones finds himself are such that it impossible
for him to do otherwise in those circumstances. Either way, the example will have failed to establish the possibility of FR-situations.¹

Defenders of Frankfurt’s argument have tried to avoid this dilemma by developing more sophisticated “Frankfurt-style” examples that either do not involve a prior-sign (e.g., Hunt 2000, Mele and Robb 1998) or in which the relation between the prior-sign and the agent’s action is such that it allows Black (or whatever plays the role of Black) to rule out all morally significant alternatives without causally determining the agent’s action (Pereboom 2000, Hunt 2005).² Whether any of the new Frankfurt-style examples succeed in avoiding the dilemma is still a matter of controversy, and no consensus on the matter has yet emerged.³

Suppose no consensus on the possibility of FR-situations can be reached. Would the debate over PAP then end in a stalemate? I should think not. Although the question of whether Frankfurt-style examples succeed in establishing FR is an important one, the intense focus on these examples has tended to obscure larger dialectical currents in the debate over PAP. In particular, it has eclipsed the question of whether agents in FR-situations could be blameworthy for their actions, as Frankfurt claims. Many participants on both sides of the debate seem to have simply taken it for granted that M, which readers will recall states that it is possible for someone in an FR-situation to be blameworthy for what he did (despite the fact that he couldn’t have avoid doing it) is true, but as we shall see, the matter is not beyond dispute. And, if it could be shown that M is false, the PAP-defender would then be able to show that Frankfurt’s argument fails quite independently of whether FR-situations are possible. This is significant, especially in light of the fact that the debate over the possibility of FR-situations seems to be moving towards a stalemate, if it has not reached that point already. So let us turn our attention to M.

¹ Kane 1985, Widerker 1995, and Ginet 1996 have each developed versions of this dilemma.
² It is generally agreed by all participants in the debate that no non-question begging example can be constructed that does not involve at least some alternative possibilities. The question, then, becomes whether the alternatives that remain in the new Frankfurt-style examples are morally significant, where a morally significant alternative (sometimes also called a ‘robust alternative’) is the sort of alternative that a defender of PAP would claim transforms a case from one in which the agent is not responsible to one in which the agent is responsible, and does so in such a way that the addition of the alternative would aid in explaining or “grounding” the claim that the agent is morally responsible.
³ For the record, I am presently agnostic about whether any of the new examples succeed in avoiding the dilemma, since in the end I think that the issue will depend on whether a certain sort of causal preemption (often called ‘trumping preemption’) is possible, and at present, I am agnostic about whether this sort of causal preemption is possible.
2.2 Two Arguments

In order to assess M there are several issues that need to be taken into consideration. Contrary to what much discussion of Frankfurt’s argument might suggest, he did not simply rely on the intuitive force of his examples to support the claim that agents in FR-situations can be responsible for their actions despite their inability to do otherwise. As Michael McKenna (2008) explains, he also offered two arguments in support of this claim. McKenna contends that these arguments rely on the following two principles, respectively:

**Irrelevance Principle**: Facts that are irrelevant to explaining why an agent performs a certain action have no bearing on whether the person is morally responsible for performing that action.

**Good Excuse Principle**: A person can cite something as an excuse only if it reveals either that the person did not act from a culpable motive, or that the person’s action issued in a causally deviant way.

Because his inability to do otherwise is supposedly irrelevant to an account of why Jones killed Smith—he would have killed him even if Black had not been present, that is, even if he could have done otherwise—it follows from the Irrelevance Principle that Jones’s inability to do otherwise has no bearing on his moral responsibility for shooting Smith. Moreover, the fact that he couldn’t have done otherwise does nothing to show that Jones did not act from a culpable motive, nor does it show that his action issued in a causally deviant fashion; the action was apparently brought about in just the way it would have been brought about had Jones been able to do otherwise. Hence, according to the Good Excuse Principle, Jones cannot cite his inability to do otherwise as an excuse for shooting Smith. So if he is not morally responsible for shooting Smith, it cannot be because he couldn’t have done otherwise but because of some other feature of the situation. Hence, whether he is responsible or not has nothing to do with his inability to do otherwise.
Unfortunately for Frankfurt and his supporters, neither argument is very persuasive, at least not when blameworthiness that is at issue. Start with the Irrelevance Principle. As it stands, this principle is false. To see this, consider an example of David Widerker’s. Imagine that “Green harms another person for some selfish reason, knowing very well that in doing so he is acting immorally.” As Widerker explains, the fact that Green knew his action was wrong arguably has a bearing on whether he is blameworthy for performing it, but this fact is irrelevant to an explanation of why Green did what he did. He did not harm the other person because he knew it was wrong to do so, but rather despite knowing this (2003, p. 62).

In a response to Widerker, Frankfurt allows that Green’s awareness that the action he performed was immoral “does not help to explain his action either by specifying one of its causes or by identifying a reason or motive for which he acted. So far as those modes of explanation go,” Frankfurt says, “Green’s knowledge that he was acting immorally plays no role in explaining his action.” Nevertheless, his knowledge “formed part of his understanding of what he was doing,” and “did not dissuade him from acting in pursuit of his selfish interests,” a fact Frankfurt insists is relevant to what he calls a “morally pertinent explanation” of Green’s action. He writes:

What counts in the assessment of a person’s moral responsibility is not only what causes, reasons, or motives led to his action. It is also important to appreciate what sort of act he thought he was performing. A morally pertinent explanation of what a person has done must include an account of what he believed himself to be doing (2003: 342-343).

On Frankfurt’s view, then, Green’s awareness that he was acting immorally is relevant to an explanation of his action after all.

We need to take a closer look at Frankfurt’s response to Widerker. Frankfurt concedes that Green’s awareness that he is acting immorally is irrelevant to what we might call a standard explanation of his action, that is, an explanation of his action in terms of its causes and his reasons for performing it. However, Frankfurt insists that Green’s awareness of the moral impermissibility of his action is nevertheless relevant to
what Frankfurt calls a *morally pertinent* explanation of that action. But what sort of explanation might that be? It seems that what Frankfurt has in mind is something like this: an explanation of a person’s action is morally pertinent provided that it includes all the facts necessary for an adequate assessment of the agent’s moral responsibility.\(^4\)

According to Frankfurt, then, the fact that Green was aware he was acting immorally, while perhaps irrelevant to a standard explanation of his action, is nevertheless relevant to a morally pertinent explanation of his action, in which case Widerker’s example is not a counterexample to a suitably refined version of the Irrelevance Principle.

It will be helpful to think of Frankfurt’s response to Widerker in the following way. Frankfurt is willing to concede that Widerker has produced a counterexample to

IRP\(_1\): When a fact is irrelevant to a *standard* explanation of a person’s action, that fact is irrelevant to whether the person is blameworthy for what he has done.

But he denies that it is a counterexample to

IRP\(_2\): When a fact is irrelevant to a *morally pertinent* explanation of a person’s action, that fact is irrelevant to whether the person is blameworthy for what he has done.

This suggests that it is the latter version of the Irrelevance Principle that is supposed to figure in the Irrelevance Argument. Therein lies the rub. In order to use IRP\(_2\) to support his claim that, in SHOOT SMITH, Jones is blameworthy for shooting Smith even though he couldn’t avoid doing so, Frankfurt would also have to assume that the fact that Jones couldn’t avoid shooting Smith is irrelevant to a morally pertinent explanation of Jones’s action. But to make that assumption, without any supporting argument, would be to beg the very question at issue. So as it stands, the Irrelevance Argument adds little, if anything, to Frankfurt’s overall case against PAP.

\(^4\) It strikes me as misleading to characterize this as an *explanation* of a person’s action, but I shall leave this worry aside.
Consider, next, the Good Excuse Principle. Here the main problem is one of understanding. How are we to take this principle? What is a culpable motive, and what, exactly, is it for an action to be produced in a causally deviant way? The present formulation of the principle is due to McKenna (2008), who derives it from some remarks of Frankfurt’s (1969), and McKenna does not address these questions in any detail. This is unfortunate, for the devil really is in the details. But even on an intuitive understanding of the principle, it isn’t clear what it has going for it. If things that have no effect on a person’s body, mind or immediate environment can be among the factors that determine whether it is within that person’s power to avoid doing what he in fact does, why couldn’t such things be among the factors that determine whether the person is morally responsible for his actions? The answer isn’t clear.

So neither of Frankfurt’s two arguments for M is terribly compelling. Is there anything else that can be said for or against M? According to the W-defender there is. In the remainder of this chapter I focus on the W-defense, which purports to give reasons why agents in FR-situations could never be blameworthy for what they have done; that is, it purports to show that M is false. The W-defense is perhaps one of the most significant responses to Frankfurt in recent years, not only because it suggests a way of advancing the debate over PAP independently of the controversy surrounding FR, but as I will explain in more detail shortly, it also raises interesting questions about the nature of moral blame. Indeed, it is an assumption about the nature of blame to which the W-defender is seemingly committed that I contend makes W-defense ultimately untenable.

2.3 Blameworthiness and the W-Defense

For the sake of argument, let us grant that SHOOT SMITH is an FR-situation: Jones is unable to avoid shooting Smith, and the circumstances that prevent him from doing otherwise do not bring it about that he shoots Smith. Having granted this, we must now ask whether Jones is morally blameworthy for what he did. Widerker claims that he is not. In support of this claim he poses the following challenge to Frankfurt and other would-be defenders of Frankfurt’s argument:

…since you…wish to hold Jones blameworthy for [shooting Smith], tell me *what, in your opinion, should he have done instead?* Now you cannot claim that he should not have [shot Smith], since this was something that was not in his power to do. Hence, I do not see how you can hold him blameworthy for [shooting him] (Widerker 2003, p. 63).

Widerker is confident that no reasonable answer can be given to the question “What should Jones have done instead?” in which case he thinks we cannot legitimately hold Jones blameworthy, for as he goes on to explain,

When we consider someone morally blameworthy for a certain act, we do so because we believe that morally speaking he should *not* have done what he did. This belief is essential to our moral disapproval of his behavior. Sometimes, however, such a belief may be unreasonable, for example, in a situation in which it is clear that the agent could not have avoided acting as he did (2003, p. 63).

According to Widerker, then, to hold Jones blameworthy would require that we believe he should have done otherwise, and to believe that a person should have done otherwise is unreasonable if the person is not able to do otherwise. Therefore, we cannot reasonably hold Jones blameworthy for what he has done even if, as Frankfurt claims, he acted of his own free will.

It is not immediately clear why Widerker deems it unreasonable to believe that an agent should not have performed some action if, given the circumstances, the agent could not have avoided performing that action. One might appeal here to the maxim that ‘ought’ implies ‘can’ (henceforth, “the Maxim”). In cases in which an agent could not have done otherwise, the Maxim implies that the agent was not obligated to do otherwise, and if an agent was not obligated to do otherwise, it is hard to see how one could reasonably claim that the agent *should* have done otherwise. However, Widerker cannot appeal to the Maxim here, for (in response to an objection raised by John Martin Fischer which I discuss shortly) he denies that the W-defense turns on the Maxim (2005 pp. 303-
What grounds Widerker’s claim, rather, seems to be this: to believe that someone should not do something involves an expectation or demand that the person not do that thing. But in a situation in which it is clear to us that the agent could not have avoided performing some action, it would be unreasonable to expect the person not to perform that action, since to do so would be to expect the person to do what for her, in those circumstances, is impossible. Hence, it would be unreasonable to believe that a person should not have done something she couldn’t have avoided doing.

According to Widerker, the W-defense thus suggests the following constraint on moral blameworthiness, which he calls the Principle of Alternative Expectations:

\[
\text{PAE: An agent } S \text{ is morally blameworthy for performing an action } A \text{ only if under the circumstances it would be morally reasonable to expect } S \text{ not to } A.\]

On the assumption that it would be morally unreasonable to expect an agent not to do something he is unable to avoid doing, PAE entails PAP.

To avoid confusion, we should distinguish between two different senses of the term ‘expect’. Consider a statement like “I expect the storm to soon pass.” In statements of this sort, the term is being used in an altogether different sense than it is being used in the W-defense or in PAE. To expect that the storm will soon pass is to anticipate or predict that the storm will be over with shortly. To expect that a person do something in the sense at issue in the W-defense, however, is to demand, insist or require that the person do that thing, as, for example, when one expects one’s children to behave in public or when one expects one’s students not to cheat on their assignments.

So Widerker’s claim, if I have understood him correctly, is that while the agent in FR-situations like Shoot Smith has perhaps violated the demands of morality (even though he couldn’t avoid violating these demands), it is unreasonable to hold him blameworthy for his wrongdoing, for in doing so we would be placing an unreasonable expectation or demand on him. By holding him responsible (e.g., by blaming him) we would be expressing an expectation or demand that he not have done something he was

---

6 Widerker goes on to embrace an account of blameworthiness based on PAE according to which an agent is blameworthy for what he did if and only if he acted wrongly, and it would be reasonable to expect him not to have acted in that way (2003, p. 68).
unable to avoid doing, and such an expectation, Widerker thinks, would be unreasonable and thus unjustified.

Before turning to some recent responses to the W-defense, I should point out an initial problem with the argument as Widerker formulates it. Contrary to what he claims, the W-defense does not directly support either PAE or PAP. To see this, notice that the conclusion of the W-defense is that we cannot legitimately hold an agent in FR-situations blameworthy. That it would be unreasonable to hold a person blameworthy, however, needn’t imply that the person is not in fact blameworthy, for a person might be blameworthy for what he has done even if it would be unreasonable for others to hold him blameworthy. For instance, there might be cases in which a person is blameworthy but in which the available evidence suggests otherwise. In such cases it would be unreasonable for others to hold the person blameworthy, given that their evidence suggests that the person is in fact blameless, and yet the person would be blameworthy nonetheless.

We can therefore distinguish between the conditions for blameworthiness, on the one hand, and the conditions for justifiably blaming (or holding blameworthy), on the other. There may be any number of considerations that make holding someone blameworthy inappropriate that in no way affects the person’s blameworthiness. So from the fact that it would be inappropriate to blame someone, we cannot conclude that the person is not blameworthy. Holding a person responsible, to put the point in a slightly different way, is not the same as the person actually being responsible. A person might be held responsible for what he has done even though he is not in fact responsible, and he might be responsible even if no one actually holds him responsible or is justified in holding him responsible.

Even if it were successful, then, the W-defense would not directly suggest PAE but rather

\[ \text{PAE}^* : \text{We can legitimately hold } S \text{ morally blameworthy for } A \text{-ing only if under the circumstances it would be morally reasonable to expect } S \text{ not to } A. \]
Unlike PAE, however, PAE* does not entail PAP, since PAP is a principle about the conditions under which an agent actually is blameworthy and not the conditions under which it would appropriate to hold an agent blameworthy. So if the W-defense is to yield a defense of PAP for blameworthiness, Widerker would first need to show that the fact that it is unreasonable to hold an agent in FR-situations blameworthy is reason to think that the agent is in fact not blameworthy.

This difficulty notwithstanding, some might argue that the W-defense provides intuitive support for the claim that agents in FR-situations should be absolved. After all, perhaps the best explanation of why it would be unreasonable to hold the agent responsible is because he is in fact not responsible. Is there anything opponents of PAP can say in response to this worry? I think there is. I shall offer my own rejoinder to the W-defense shortly, but before doing so I first examine three recent replies to the W-defense, which for various reasons I deem problematic.

2.4 Some Unsuccessful Replies

Eleonore Stump (2003) suggests that if we pay careful attention to the way in which the agent’s action actually comes about in FR-scenarios and the way it would have come about had the intervener been forced to take action, we will discover an adequate response to the question “what should Jones have done instead?” In an example like Shoot Smith, although it is not up to the agent whether he performs a certain action, it is up to him, Stump says, how the action he performs is produced. More precisely, it is up to him whether his action is a result of his own agency or whether it is the result of Black’s intervention. The right answer to Widerker’s question, then, is that Jones should have brought it about that his action is a result of Black’s intervention and not of his (Jones’s) own free agency.7

This sort of reply is problematic, however, for a PAP-defender might insist that while perhaps providing a satisfactory response to the question at hand, the reply does so

7 A similar response, which was suggested to me by Al Mele, is that Jones should not have shot Smith freely. What he should have done instead is shoot Smith unfreely; that is, he should have brought it about that he is forced to shoot Smith. Since this is something he could have done, it is a reasonable answer to Widerker’s question. The same sorts of worries I raise for Stump’s response apply, mutatis mutandis, to this response.
only by pointing to the fact that there are at least some alternative possibilities available to Jones after all, which is precisely what many defenders of PAP have said about FR-situations all along. Stump’s response thus avoids the force of the W-defense (if it does) only to fall prey to a different sort of objection to Frankfurt’s argument, one according to which the Frankfurt-style examples do not rule out all morally significant alternative possibilities. This objection has been called the flicker of freedom defense.8

Stump anticipates this worry and attempts to forestall it. She says that although the agent does have alternative possibilities of some sort, they are not alternative actions. That the agent has control over how his action is produced is not to say that he has control over which action he performs. The agent may have no choice but to perform a certain action while nevertheless retaining control over how that action comes about. PAP-defenders therefore cannot appeal to such alternatives in support of their principle, since PAP is a principle requiring alternative courses of action and not alternative possibilities more generally.

This response seems to me to miss a larger point. Even if what Stump says is correct (and it is, as far as it goes), it would not resolve the larger debate between advocates of the flicker of freedom strategy and Frankfurt-defenders. Advocates of the flicker of freedom defense typically are not committed to PAP per se but rather to an account of freedom and responsibility according to which at least some alternative possibilities are required for moral responsibility. PAP provides a natural expression of this view, but should that principle be false, for precisely the sorts of reasons Frankfurt adduces, this would not imply that the alternative possibilities approach to responsibility is mistaken, for as a number of people have pointed out, there may be principles other than PAP which have the consequence that moral responsibility requires alternative possibilities, principles that are not threatened by Frankfurt’s case against PAP.9 Many of those who have defended Frankfurt’s argument, however, including Frankfurt himself, have done so in service of what have come to be known as actual sequence accounts of freedom and responsibility. According to actual sequence theorists, what is important

---

8 The flicker of freedom strategy is developed in different ways by van Inwagen 1978, Naylor 1984, Otsuka 1998, and Wyma 1997. The term flicker of freedom was first coined by Fischer 1994 as a way of characterizing these sorts of responses to Frankfurt.
when it comes to moral responsibility is the action the agent actually performed and how that action was brought about. Whatever alternative courses of action, or alternative possibilities more generally, may have been available to the agent are irrelevant to whether the agent is morally responsible. So while Stump’s reply might provide the Frankfurt-defender with a technical victory in the case of PAP, her remarks leave the spirit of Frankfurt’s argument in doubt insofar as they invite proponents of the flicker defense to press their claim that there are no cases of responsibility without alternative possibilities.

Of course, it may be that the Frankfurt-defender has an adequate reply to the flicker defense. Such a reply, however, would only return us to the controversy over FR and whether a Frankfurt-style example can be constructed that precludes all morally significant alternatives without causally determining the agent’s action. Stump’s response to the W-defense therefore provides a cogent defense of Frankfurt’s argument only if the larger issues surrounding FR have an adequate, Frankfurt-friendly resolution. But part of what I hope to show is that there is a good response to the W-defense that can be given quite independently of the controversy surrounding FR. Moreover, a plausible response to the flicker of freedom defense would likely rule out Stump’s reply to the W-defense. In order to respond to the flicker defense, one would have to construct an example in which any remaining alternatives are either beyond the agent’s control or, though within the purview of the agent’s control, are morally insignificant in the sense that their presence is irrelevant to the issue of whether the agent is morally responsible. If the remaining alternatives were beyond the agent’s control, it would be unreasonable to expect the agent to avail herself of those alternatives. If, on the other hand, the remaining alternatives were morally insignificant, this would seem to contradict the assumption implicit in Stump’s remarks that the alternatives to which she claims the agent has access are relevant to whether the agent is morally responsible. In either case, then, Stump’s reply to the W-defense would no longer be available.

Recently, Frankfurt himself has returned to the fray and has offered a reply to the W-defense. Here is the key part of what he has to say about the matter:
What should Jones have done instead of what he did? He should have behaved virtuously. In deciding of his own free will to [shoot Smith], he violated the requirements of morality. Instead, of doing that, he should have refrained from violating those requirements (2003, p. 343).

Frankfurt’s remarks suggest that he is interpreting the question “what should Jones have done?” as a question about what Jones was morally obligated to do. As I explain below, this is apparently not how Widerker intends the question to be understood, but before I do so, it will prove instructive to first examine Frankfurt’s answer to the question as he (Frankfurt) seems to be interpreting it.

There are at least two ways to read Frankfurt’s remarks.\footnote{I am grateful to Al Mele for pointing me to this first reading.} Frankfurt says that “in deciding of his own free will” to shoot Smith, Jones violated the requirements of morality. This might suggest that on Frankfurt’s view Jones violated the demands of morality not by shooting Smith \textit{per se} but rather by shooting Smith \textit{of his own free will}. So what Jones should have done, according to this first reading, is to refrain from shooting Smith of his own free will. The virtue of this reading is that it allows Frankfurt to retain the Maxim that ‘ought’ implies ‘can’, since Jones could have refrained from shooting Smith of his own free will by bringing it about that he is forced by Black to shoot Smith. However, this reading of Frankfurt’s response is simply a variation of Stump’s response and is subject to the same sorts of worries.

A second reading is that Jones violated the demands of morality simply by shooting Smith (full stop). Thus, what he should have done is to refrain from shooting Smith. The virtue of this reading is that it distinguishes Frankfurt’s response from that of Stump. The (potential) downside is it also seems to require that Frankfurt abandon the Maxim that ‘ought’ implies ‘can’, for on this second reading, Jones should have refrained from shooting Smith even though it was not within his power to refrain from shooting him. Some might find this second reading problematic for precisely this reason, for the Maxim is regarded by many as a highly plausible principle. As Fischer (2006, pp. 203-209) has pointed out, however, the Frankfurt-defender need not find this problematic; it is only natural for those sympathetic with Frankfurt’s argument to reject the Maxim along
with PAP. Indeed, many actual sequence theorists would point to Frankfurt-style examples as counterexamples to both PAP and the Maxim. By shooting Smith it seems that Jones has violated an obligation not to murder (or some similar obligation, e.g. not to harm others needlessly), an obligation he would not have if the Maxim were true. Fischer contends, moreover, that rejecting the Maxim “completely disarms the W-defense” (2006, pp. 210). Fischer’s idea seems to be that without the Maxim there is no longer any barrier to saying that Jones should have done otherwise.

There is reason to believe that both Frankfurt (on this second reading) and Fischer have missed their target, for as I noted earlier, Widerker insists that the W-defense does not depend on the Maxim. He says there may be circumstances in which an agent is morally obligated to behave in a certain way but in which it would be unreasonable to expect the person to comply with the demands of morality precisely because the person is unable to do so. On his view, then, it appears that Jones may well be morally obligated to refrain from shooting Smith even though it is not within his power to refrain from shooting him. Nevertheless, according to Widerker, it would still be unreasonable to expect Jones to fulfill his obligation not to shoot Smith given that it is an obligation with which he is unable to comply. Denying the Maxim therefore apparently does not disarm the W-defense.\footnote{Fischer (personal correspondence) has suggested to me that, while perhaps not “completely disarming” the W-defense, denying the Maxim might make the W-defense substantially less plausible, for why should one find either PAE or PAE* plausible if one has jettisoned the Maxim? I concede that there might be something to this suggestion, but notice too that there might be reasons for accepting PAE or PAE* even if the Maxim is false. For instance, as I suggest later, in contexts in which one makes judgments about what a person morally ought to do, there is often an accompanying expectation or demand that the person behave as she ought. It is arguably part of our moral practices that we demand of others that they behave as they ought, but such demands seem reasonable only in cases in which people can behave as they ought. Thus, one might find the underlying principles behind the W-defense plausible even if one denies that being morally obligated to act in a certain way implies that one is able to act in that way.}

Widerker’s response to Fischer suggests that the question “what should Jones have done?” is not to be understood primarily as a question about what Jones is morally obligated to do but rather as a question about what it would be reasonable to expect him to have done given the circumstances. Thus neither Frankfurt nor Fischer’s reply to the W-defense meets the challenge set out by Widerker, for neither reply provides an answer to Widerker’s question, thus construed. An adequate answer to the challenge posed by the W-defender would need to say more than just what the agent was morally obligated to do
or not do. It would also need to say what it would be reasonable to expect the agent to do, given the circumstances. If no answer is forthcoming, then, on Widerker’s view, we cannot legitimately hold an agent in Frankfurt-style cases blameworthy.

In my estimation no reasonable answer to the question “what should Jones have done?” can be given. Where I take issue with the W-defense is the assumption that if no reasonable answer is forthcoming, then we should let agents in FR-situations off the hook. As I shall argue shortly, the Frankfurt defender simply has no reason to accept this assumption. However, before doing so, it will prove instructive to look at one last reply to the W-defense on the part of McKenna (2005 and 2008).

McKenna concedes that the W-defense cannot be decisively refuted. He and I agree that there is no reasonable answer to Widerker’s challenge that does not, in one way or another, play into the PAP-defender’s hands. However, McKenna seems to endorse the assumption implicit in the W-defense that holding an agent blameworthy involves an expectation that the agent not behave as she did. It is worth quoting him at some length on this matter:

> In a context regarding judgments of blame, in stating what a person should have done but did not do, there is some presumption that she could have done that thing.

> I think that Widerker’s W-Defense cannot be directly refuted; there is something right about it. Widerker has tapped into an important moral intuition favoring a conception of free agency and moral responsibility in terms of alternative possibilities. Some arguing for Frankfurt’s thesis are reluctant to admit that there could remain residual intuitions of our moral thought that cannot be fully accommodated. But Widerker’s W-Defense has simply hit one of those intuitions spot on. It is only philosophically honest to acknowledge that…. My assessment of the W-Defense is that it is such an impressive argument because it hits upon a raw nerve for the Frankfurt Defender. By way of the W-Defense, certain features of ordinary thought are brought into relief; Frankfurt’s

---

12 One response to the W-defense that I do not discuss is Zimmerman 2003. He raises some interesting questions about what precisely Widerker means when he speaks of morally reasonable expectations. In the present essay I have been content to rely on an intuitive understanding of this notion.
conclusions do lead to an “irresolvable tension” in our thinking about blame and its implications (regarding what an agent should have done instead). (2008, p. 784)

Despite his willingness to concede the force of the W-defense, McKenna goes on to argue that its force can be counterbalanced by what he calls the L-reply.

McKenna is an actual sequence theorist. Consequently, he insists that what we should be concerned with is the actual sequence of events. When it comes to freedom and responsibility what matters is not what the agent could or could not have done but what the agent did and how the agent’s action was actually produced. So in response to Widerker’s question “What should Jones have done instead” McKenna offers the L-reply: “Look at what Jones has done.” The L-reply forces us to look not at alternative possibilities or the lack thereof but on the action the agent actually performed and the moral quality of the agent’s will that is revealed in that action. By forcing us to look at these features of the situation, the L-reply points to a tension in the thinking of the PAP-defender: those wishing to absolve someone like Jones are forced to deny the intuition that agents who freely and deliberately perform actions they know to be morally wrong, and who thereby display an objectionable quality of will, are blameworthy. In this way McKenna thinks he can defuse the force of the W-defense by showing that, although the W-defender has pointed to an irresolvable tension in Frankfurt-inspired actual-sequence views, the W-defender is committed to a different, though equally irresolvable, tension in his moral thinking when he absolves agents in FR-situations of responsibility. Both views, then, leave us with residual intuitions that we cannot fully explain away. But then the W-defense is not itself sufficient to refute M. Both those who accept and those who reject M will have to appeal to further considerations to support their respective views about the blameworthiness of agents in FR-situations.

The virtue of the L-reply is that once we focus on the actual sequence of events and the moral quality of the agent’s will that is revealed in his action, it does seem initially plausible to suppose that agents in FR-situations who perform morally wrong actions are blameworthy for what they have done. The difficulty with McKenna’s reply, as I see it, is that it concedes too much to the W-defender. Those wishing to defend
Frankfurt’s thesis, I shall argue, need not admit that their position leaves them with an “irresolvable tension” in their thinking about blame and its implications. I think further examination of the W-defense will reveal that the moral intuition which that argument “taps into,” as McKenna puts it, is not as plausible as it might first appear.

2.5 A Rejoinder to the W-Defense

Having explained why I find some of the previously published responses to the W-defense unsatisfying, I now develop a rejoinder of my own. As we have seen, it is the W-defender’s contention that holding a person blameworthy for her wrongdoing essentially involves an expectation or demand on our part that the person not have behaved as she did. But the W-defender insists that such an expectation would be unreasonable if the agent was unable to avoid acting as she did. Thus, in situations in which a person could not have avoided performing some action, we cannot reasonably blame that person for what she has done. This argument can be captured in the following schema:

1. To hold S blameworthy for A-ing essentially involves an expectation that S not have A-ed (premise).
2. But to expect that S not have A-ed would be unreasonable if S could not have avoided A-ing (premise).
3. Hence, holding S blameworthy for A-ing when S could not have avoided A-ing would involve an unreasonable expectation on the part of those holding S blameworthy (from 1 and 2).

I take it that 3 constitutes the sort of “irresolvable tension” in Frankfurt’s thought that McKenna thinks the W-defense brings into relief. It points to an apparent unreasonableness to which one would have to be committed if one were to embrace Frankfurt’s conclusions about the blameworthiness of agents in FR-situations.

How cogent is the W-defense thus construed? Not very; both premises are suspect. Let me start by noting a problem with premise 2. I will then address premise 1, which I see as the main claim underlying the W-defense.
Suppose for the moment that 1 is true. Even so, 2 is arguably false. To see this, consider the following scenario. Kevin has an opportunity to steal my wallet. Suppose I expect Kevin, who as far as I know is a normal human adult, not to steal my wallet, despite the fact that he has every opportunity to steal it. Unbeknownst to me, however, Kevin is kleptomaniac who literally cannot help stealing when the opportunity to steal presents itself. If 2 were true, it would follow that I am being unreasonable in expecting Kevin not to steal from me. But that seems implausible. Given what I know about Kevin, my expectation that he not steal my wallet seems perfectly reasonable, even though he cannot help stealing it. Hence, 2 is false.

In light of examples like this, W-defenders might try revising the second premise of their argument as follows:

2’. To expect that S not have A-ed would be unreasonable if one knew (or perhaps reasonably suspected) that S could not have avoided A-ing.

This claim is perhaps plausible, certainly more plausible than 2. Notice, however, that 3 does not follow from 1 and 2’. What follows, rather, is

3’. Holding S blameworthy for A-ing despite knowing (or reasonably suspecting) that S could not have avoided A-ing would involve an unreasonable expectation on the part of those holding S’ blameworthy.

This is a significantly weaker conclusion than 3. Thus, at most the W-defense shows that our practices of holding one another blameworthy would be unreasonable in those cases in which we are aware (or at least reasonably suspect) that the agent couldn’t have avoided doing what he did.

Having noted this limitation of the W-defense, I turn now to premise 1, which, as I said, I take to be the driving claim behind the Widerker’s argument. Why think that blaming a person for what she has done essentially involves an expectation or demand that she not have done what it is we are blaming her for doing? The answer is not at all obvious. Perhaps one reason some might find 1 appealing has to do with the fact that in
paradigm cases of blameworthiness it is plausible to suppose that the agent could have done otherwise. It is, I think, part of our commonsense picture of ourselves as free agents that we can often do otherwise than what we in fact do. So in normal cases in which it is appropriate to blame a person for her actions, it might well be reasonable to expect that the person have done otherwise. It might even be the case that in most instances there is an expectation that the person have done otherwise. But clearly this need not imply that all instances of holding someone blameworthy must involve an expectation that the person have acted differently. Perhaps there are non-paradigmatic cases in which we may reasonably blame someone for what she has done even though we do not (and could not reasonably) expect her to do otherwise.

In his original essay on PAP, Frankfurt pointed out that that principle gains plausibility from the fact that when considering cases in which a person could not have done otherwise, we tend to think only of those cases in which the circumstances that prevent the person from doing otherwise also play a role in producing the agent’s action. Frankfurt suggests, however, that if we focus on atypical, non-paradigmatic cases, namely, FR-situations, it becomes significantly less clear that PAP is true. Focusing only on the paradigmatic instances of “couldn’t have done otherwise,” cases in which the agent is forced, coerced, or otherwise compelled against her will, does indeed suggest something like PAP. It is only by looking at the non-paradigmatic cases, however, that we can discover potential counterexamples to the principle.

A similar point applies to premise 1. If we focus only on paradigmatic cases of holding blameworthy, it might seem plausible to suppose that all cases of holding blameworthy involve an expectation that the person have done otherwise, for as I have suggested, in those paradigmatic cases we may typically think it is reasonable to expect the person to have done otherwise—because we believe the person could have done otherwise—and we may in fact expect the person to have done otherwise. But in order to assess the plausibility of 1, we need to look at atypical cases as well. Once we do so, however, it is less clear that holding someone blameworthy essentially involves an expectation that the person have acted differently. As McKenna rightly points out, when we focus on the actual sequence of events in Frankfurt-style cases, and especially the fact that agents in those situations act entirely on their own accord, there is pressure to say
that the agent is blameworthy for what he did despite the fact that we could not reasonably expect him to do otherwise. So in the absence of further reason to think that holding a person blameworthy essentially involves an expectation that the person have done otherwise, those who share these intuitions about the blameworthiness of agents in FR-situations are well within their rights to deny 1 in favor of M.

A second reason some might find premise 1 initially plausible has to do with the fact that holding a person blameworthy invariably involves a conative element. When we blame someone for what she has done, we desire that the person not have done what she did. This desire, moreover, is arguably essential to holding a person blameworthy. To see this, imagine that the devil is evaluating Jones’s behavior. The devil might agree with Frankfurt at the cognitive level that Jones is blame-worthy—i.e., that he meets the relevant conditions for moral responsibility—but such a cognitive judgment on the part of the devil does not mean that the devil actually holds Jones blameworthy. After all, the devil is no doubt gleeful that Jones has done the morally wrong thing. Indeed, we may suppose that it was the devil who sent Black to insure that Jones shoot Smith! In order for the devil to actually hold Jones blameworthy, it seems that there needs to be a conative element present as well. In order to actually blame Jones, the devil must also desire that Jones not have shot Smith.

So blame involves both a cognitive and a conative element, and it is perhaps not difficult to see how a desire that S not A, which is essential to holding S blameworthy for A-ing, might be confused with, or seen as involving or somehow being associated with, an expectation or demand that S not A. Typically, when we expect a person not to do something, our expectation is accompanied by a desire that the person not do that thing. It is important to notice, however, that to desire that S not A needn’t point to the presence of an expectation or demand that S not A. I might desire that you perform a certain action even if I cannot reasonably expect you to perform that action. I might desire that you give me a million dollars, for example, but it would clearly be unreasonable for me to demand this of you. Or, to return to FR-situations, we might desire that Jones not shoot Smith, but given the circumstances it would be unreasonable to expect Jones not to shoot Smith. The fact that holding a person blameworthy involves a conative element therefore does not support premise 1 of the W-defense either.
Widerker claims that an expectation that a person not act as she did is essential to our moral disapproval of her behavior. The Frankfurt-defender, I am suggesting, has been given no reason to accept this claim. Why must we expect a person to do otherwise in order to disapprove morally of the person’s behavior? Those who side with Frankfurt will insist that it is entirely possible to disapprove of a person’s behavior and indeed to hold that person blameworthy for what she has done without there being an expectation on our part that she not have behaved in that way. To be sure, we will no doubt wish that the person not have behaved as she did, and we may judge that the person ought not have behaved in that way (thereby rejecting the ‘ought’ implies ‘can’ principle), but as I have been arguing, none of this need involve an expectation or demand on our part that the person not behave as she did.

So what should Jones have done instead of shooting Smith? Well, if we mean “what ought he to have done instead?” then a Frankfurt-defender might deny that ‘ought’ implies ‘can’ and insist that Jones should have refrained from shooting Smith. If, on the other hand, we mean “what would it be reasonable to expect (or demand) Jones to have done instead?” then the correct answer, I think, is “there is nothing we could have reasonably expected him to do instead.” Since we know that Jones could not have done otherwise than shoot Smith, it would be quite unreasonable to expect him not to shoot Smith. I have argued that this is problematic for the Frankfurt-defender only if by holding Jones blameworthy the Frankfurt-defender is thereby committed to expecting that Jones not have done what he did. But we haven’t been given any reason to suppose that the Frankfurt-defender is committed to such an expectation, for the W-defender hasn’t given us any reason to suppose that an expectation that a person not have acted in a certain way is essential to our moral disapproval of that person’s behavior or to our holding that person blameworthy for what he has done. So in the absence of any further defense of step 1 of the W-defense, I see nothing problematic about those who are inclined to reject PAP on the basis of Frankfurt-style cases rejecting 1 in favor of M.

2.6 A Tentative Suggestion

I conclude with a tentative suggestion. Up to this point, the debate over PAP has largely focused on metaphysical considerations regarding the nature of free agency. However,
my discussion of the W-defense suggests that many of the intuitions driving various positions in this debate might depend to a greater or lesser extent on particular accounts of the nature of moral responsibility and allied notions such as praise and blame. In light of this, the debate over PAP might benefit from a closer examination of these moral concepts. Whether moral responsibility requires the ability to do otherwise might depend, in the final analysis, on which account of moral responsibility one adopts. In this chapter I have argued that the account of blame implicit in the W-defense has little to recommend itself, which leads me to the conclusion that the W-defense is ultimately unsuccessful.
CHAPTER 3

ACTION, RESPONSIBILITY AND ABILITY

According to what we might call the Principle of Possible Refrainment,

PPR: An event $A$ constitutes an action on the part of an agent $S$ only if it was within $S$’s power to refrain from $A$-ing.

PPR is interesting in its own right, and has obvious implications for a theory of action. It will therefore be of interest to philosophers concerned with the nature of action generally. It will also be of interest to those concerned specifically with free action and moral responsibility. If PPR is true, then a fortiori a person performs a free action for which he is responsible only if he could have refrained from performing it. But PPR is implausible; there is little, if any, reason to think it is true, and good reasons to think it is false—or at any rate, so I shall argue. If I am right, then three recent defenses of PAP are unsuccessful.

3.1 Preliminaries

PPR has recently been defended by both Maria Alvarez (2009) and Helen Steward (2008; 2009). According to Alvarez, “for what someone does to be his action, for him to be its agent, the person must have a certain degree of control over it: at least, he should be capable of refraining from doing what he does” (2009, p. 74). Similarly, Steward claims that in order for $A$ to count as an action, it “will have to have been the product of an exercise of what is sometimes called a two-way power on the part of that agent—to act, or to refrain from acting—so that in the absence of the relevant power of refrainment a genuine action simply could not occur” (2008, p. 140). Elsewhere she says that “A (particular) action qualifies as an action…only if it is something from which the agent is able to refrain” (2009, p. 84). Moreover, both Alvarez and Steward employ the principle
in defense of PAP. It can be used to formulate an argument for PAP, and it can also be used to challenge the coherence of Frankfurt-style cases. I will take a closer look at their arguments shortly. But first, PPR requires clarification. Two issues merit attention.

The first concerns refraining. PPR says that what someone has done counts as an action of his only if he could have refrained from doing it. But what does it mean to say that the person could have refrained from doing what he did? Sometimes refraining is viewed as a type of action, or at least as involving an action. The idea is that in order to refrain from A-ing the agent must decide not to A, and deciding is a mental action. But according to Alvarez at least, refraining does not require performing a mental act of deciding. On her view, in order for it to be the case that S could have refrained from A-ing, two conditions needs to be satisfied. First, it must have been possible (she does not say in what sense) that S not A, and second, it must have been up to S whether he A-ed or not (p. 63). Thus, Alvarez’s view seems to be that S could have refrained from A-ing if and only if S could have avoided A-ing. Since the versions of PAP with which I am concerned are formulated in terms of ‘could have avoided’ (rather than ‘could have done otherwise’), for simplicity sake I will henceforth formulate PPR using this same terminology. Thus understood, the principle says that an event A is an action of S’s only if S could have avoided A-ing.

The second issue that requires attention is the sense of ‘could’ at issue in PPR. Claims about what an agent could have done are notoriously and multiply ambiguous. At the most basic level, such claims are ambiguous because they can be used to express claims about what it is possible for an agent to have done, as well as claims about what an agent was able (or had it within his power) to do. Of course, possibility claims and ability claims are not unrelated. Claims about what a person is able to do arguably imply certain claims about what it is possible (in some sense of possible) for the agent to do. But still, claims about possibility and ability are distinct.

So is PPR primarily a claim about possibility or ability? Several remarks by both Alvarez and Steward suggest that they view the principle as a claim about ability. Recall that, according to Alvarez, in order for A to be an action of S’s, S should be capable of refraining from A-ing. Similarly, recall Steward’s claim that agency requires a two-way
power and that A counts as an action of S’s only if S is able to refrain from A-ing. Following their lead, I too shall treat PPR as a claim about ability.

But what sort of ability exactly? In what sense must S have been able to avoid A-ing, according to PPR, if A is to be included among S’s actions? Here neither Alvarez nor Steward is as clear as one might have hoped. The use of terms like “capable” by Alvarez or the talk of “two-ways powers” employed by Steward might be taken to suggest that the kind of ability at issue is a general ability or capacity to avoid A-ing. At other times, however, something stronger seems to be implied. For instance, Alvarez (p. 64) suggests that for S to have the relevant ability to refrain from A-ing at t, at t, it must be up to S whether or not he A-s then. However, S might retain a general ability or capacity at t, and yet, it might not be up to S at t whether he exercises that ability or capacity at t. (Were I locked in an empty room right now, I would presumably retain lots of general abilities—e.g., the ability to type, play basketball, read a book, etc.—even though it would not presently be up to me whether I exercised those abilities just now.) This suggests that perhaps Alvarez has something stronger in mind than a general capacity. According to Steward, determinism is incompatible with the relevant sort of ability to do otherwise and is thus incompatible with genuine agency.¹ The fact that determinism is compatible with having unexercised capacities, skills or competencies suggests that Steward, too, has a stronger sense of ability in mind.

If PPR is to be used to defend PAP or similar theses relating responsibility to alternative possibilities, and in particular if PPR is to serve as a premise in an argument for PAP, ‘could have refrained’ must mean the same thing in both principles. Otherwise, any argument from PPR to PAP will be guilty of equivocation. So what does ‘could have refrained’ mean as it appears in certain formulations of PAP? It would seem to be something stronger (requiring more) than a general ability, capacity or skill, at least as the principle is understood by many of those who have contributed to this literature. If PAP were construed in terms of a general ability or capacity it would be difficult to see what all the fuss surrounding the Frankfurt-style cases has been about. In such cases, it has been claimed, the agent can be responsible for his action even though it was apparently not within his power to do otherwise. Now, whatever else is true about the agent in these

¹ See Steward 2008 in particular.
examples, it seems he retains a general capacity to do otherwise than he actually did. This suggests that Frankfurt and others take PAP to require something more than a general ability to do otherwise. Moreover, there do seem to be other abilities that the agent plausibly lacks the absence of which Frankfurt and others might insist are irrelevant to the person’s moral responsibility. For instance, although the agent may retain a general capacity to do something other than what he actually does, given the circumstances it seems fair to say that he is unable to exercise that capacity, that it is not within his power to exercise it, that it is not up to him whether he performs the action in question, and that he has no choice about whether to perform this action.\(^2\) If PAP does concern some type of ability over and above a general capacity or skill—and I shall henceforth assume that it does—then to avoid equivocation those who wish to argue from PPR to PAP should construe the former principle in such a way that it concerns this same sort of ability. Thus, claims about what an agent “could have done,” “was able to do,” or “had it within his power to do” should henceforth be read as claims about a kind of ability that is stronger than a general ability, capacity or skill.

### 3.2 Action and Responsibility

I now turn to the application of PPR to the controversy surrounding PAP. PPR is relevant to that controversy in two ways. First, it can be used to formulate an argument for (a version of) PAP, and second, it can be used to challenge the coherence of Frankfurt-style cases. I take up both issues in turn.

According to Steward, “a (particular) action qualifies as an action in the first place (and therefore as a candidate for something for which an agent might be morally responsible) only if it is something from which the agent was able to refrain (2009, p. 84). These remarks are suggestive of the following argument:

1. \(A\) is a candidate for something for which \(S\) might be morally responsible only if \(A\) is among \(S\)’s actions (premise).

\(^2\) For these points, I’m indebted to Clarke 2009, p. 340.
2. *A* constitutes is an action of *S*’s only if it was within *S*’s power to avoid *A*-ing (premise: PPR).

3. Therefore, *A* is a candidate for something for which *S* might be morally responsible only if it was within *S*’s power to avoid *A*-ing (from 1 and 2).

The conclusion of this argument is a version of PAP.

It is worth pointing out that, as it is presently formulated, the argument is unsound for reasons that have nothing whatsoever to do with PPR. Even if PPR (i.e., premise 2) is true, premise 1 is surely false. In addition to being responsible for our actions, we can also be responsible for our omissions and for the consequences of our actions and omissions.³ Actions are therefore not the only candidates for things for which we might be to blame. This defect in the argument can easily be remedied, however, by appealing to the distinction between direct and indirect responsibility introduced in chapter one.

Recall that a person is directly responsible for an event or state of affairs provided that he is responsible for it, and his responsibility for it does not depend on or derive from his responsibility for some distinct event or state of affairs. By contrast, if a person is indirectly responsible for something, his responsibility for it will depend on or be derived from his responsibility for some distinct state or event. Now, although most theories of moral responsibility allow that a person can be responsible for things other than actions, it is widely agreed that a person is directly responsible only for his actions or perhaps for some sub-class thereof, such as choices or decisions. On this fairly standard view, *S*’s responsibility for anything other than his actions depends upon its standing in an appropriate relation to an action for which *S* is morally responsible. For present purposes, I shall assume that something along these lines is correct.⁴

³ Some consequences of actions are themselves actions, but others are not. Also, some readers might deny that omissions can have consequences because they deny that omissions can be causes. But even if omissions cannot be causes, there is a perfectly ordinary and acceptable sense in which our omissions have consequences. Even if my omitting to water your flowers is not among the causes of the flowers withering, it would still be perfectly acceptable to say that the flowers withered in part because I omitted to water them. Fortunately, nothing I say here hinges on these issues. The key thing to note for present purposes is that direct responsibility is typically restricted to actions.

⁴ In fact, however, I am inclined to reject the standard view. In my opinion, a person can be directly responsible for things other than his actions, including at least some omissions. But since I think the argument could be reformulated to accommodate my non-standard view, I put these complications aside.
We can now construct an improved argument in defense of PAP based on PPR. It can be stated as follows:

1′. \( S \) is directly responsible for \( A \)-ing only if \( A \) is among \( S \)’s actions (premise).

2. \( A \) is among \( S \)’s actions only if it was within \( S \)’s power to refrain from \( A \)-ing (premise: PPR).

3′. Therefore, \( S \) is directly responsible for \( A \)-ing only if it was within \( S \)’s power to refrain from \( A \)-ing (from 1′ and 2).

The conclusion of this argument is a version of PAPd (see chapter one). Since the argument is valid, the only question is whether its second premise, which is a version of PPR, is true. I take up the question shortly.

PPR can also be used to challenge the coherence of Frankfurt-style cases. In a standard variation of Frankfurt’s original example, Jones shoots and kills Smith without hesitation. Unbeknownst to Jones, however, the evil genius Black has the ability to both read and control Jones’s mind with a sophisticated brain-scan device. Had Jones not killed Smith on his own, Black was prepared to use this device to send a signal to Jones’s brain that would have caused Jones to (decide to) kill Smith. It seems that in this example Jones is unable to avoid killing Smith. Evidently, his only options are to kill Smith on his own or be forced to kill Smith. Either way, he kills Smith. But since Jones kills Smith on his own, without any “help” from Black, Frankfurt and others believe that Jones bears just as much responsibility for what he has done as he would have had he been able to do otherwise. Thus, they conclude, PAP is false.\(^5\) This sort of argument cannot succeed, however, if PPR is true, since any “successful” counterexample to PAP would involve a conceptual impossibility, namely, a case in which a person performs an action from which he cannot refrain.

Those who believe that Frankfurt-style cases are counterexamples to PAP might try to turn this sort of objection on its head. Consider a version of the example \textit{sans} Black and his device. Call this the \textit{truncated story}. Barring any special assumptions, it seems

\(^{5}\) For a good overview of the recent literature on Frankfurt-style cases, see the introduction to Widerker and McKenna 2003.
evident that in the truncated story Jones’s killing Smith constitutes an action of Jones’s for which he can be held accountable. But if we are willing to count Jones’s killing Smith as an action in the truncated story, shouldn’t we also count it as an action of his in the original, Frankfurt-style version of the story? After all, in neither version do Black or his device play a role in the etiology of what Jones does; it seems that the sequence of events leading up to and including the actual killing of Smith is exactly the same in both versions of story. For this reason, isn’t it plausible to suppose that Jones performs an action—killing Smith—in both the original story and in the truncated story? In the Frankfurt-style case, however, Jones is apparently unable to avoid performing this action. Thus, one might contend, Frankfurt-style cases themselves provide us with counterexamples to PPR.

John Martin Fischer (2008, p. 170-172) suggests an objection to PPR along these lines. On Fischer’s view, the notion of an action is an actual sequence notion; whether \( A \) counts as an action depends solely on what happens in the actual sequence of positive events leading up to and including \( A \) and not on whether the agent has access to some counterfactual sequence of events in which \( A \) does not occur.\(^6\) To motivate an actual sequence view of action, Fischer asks us to “compare Jones in an ordinary context with Jones in a Frankfurt-type example.” He points out that, “In both scenarios there is no intervention by an individual such as Black (or any similar factor); in both scenarios the actual sequence flows in the same way.” He then insists that “if we are willing to say that Jones acts in the ordinary case…we should also say that he acts in the Frankfurt-style case,” since “everything that happens in the actual sequence is exactly the same as in the ordinary case!” (p. 170). Let us agree with Fischer that Jones does indeed perform an action in both cases. If, as Fischer thinks, it is also true that in the Frankfurt-style version of the story Jones cannot refrain from performing this action, it follows that PPR is false.

This sort of objection to PPR is not without merit. Indeed, I’m inclined to agree with Fischer that Frankfurt-style cases are counterexamples to PPR. Presumably, however, proponents of PPR will demur. According to Alvarez, for example, if Frankfurt-style cases are genuine counterexamples to the likes of PAP and PPR, then in

\(^6\) Fischer 1982 first used the term ‘actual sequence’ to describe theories of free will and moral responsibility according to which freedom and responsibility depend on the nature of the actual sequence of events leading up to the action and not on whether the agent has access to alternative possibilities for action.
the counterfactual sequence of events in which Black intervenes and forces Jones’s hand it must be the case that Black causes Jones to perform the same action (or at least an action of the same type) that Jones performs in the actual sequence of events. However, Alvarez insists that “it is not legitimate for a Frankfurt-style case simply to stipulate that, in the counterfactual case, the agent would be caused to perform an action that he cannot avoid performing…Rather, any example needs to tell a compelling story that makes the suggestion plausible without begging the issues at hand (op. cit. p. 67). But she is skeptical about whether such a story can be told, “because it is far from clear that the idea that a person can be caused to perform an action in such a way that he could not have avoided performing it but which is nonetheless his action is cogent” (op. cit. p. 72).

Steward expresses similar skepticism. She says that

It is very plausible that when Black intervenes, what he produces cannot be an event identical with the action that Jones actually undertakes when he freely [shoots Smith]. Indeed, it might be argued that Jones has it in his power not to perform an action of [shooting Smith] at all—for one might reasonably ask what reason there is to suppose that a chain of events which is initiated by Black might count…as an action on the part of Jones. (2008, p. 139)

There are several ways those wishing to defend the coherence of Frankfurt-style cases might respond to Alvarez’s worries. Fischer rejects her assumption that in order for Frankfurt-style cases to succeed Black must cause Jones to perform the same action in the counterfactual sequence that he performs in the actual sequence. On Fischer’s view, Frankfurt-style cases successfully cast doubt on principles like PAP even if the agent in the example does not perform the same action in the counterfactual sequence that he performs in the actual sequence, and, indeed, even if the agent does not perform an action at all in the counterfactual sequence. Alternatively, one might take up Alvarez’s challenge to tell a “compelling story” in which it seems plausible to say that the agent is caused to perform an action in such a way that he cannot avoid performing it. Can such a story be told? I believe so, and if I am right, then PPR is false. However, I shall not attempt to tell such a story here. Instead, I shall highlight some reasons to be suspicious
of PPR and the sorts of considerations motivating it that are independent of the challenges posed for the principle by Frankfurt-style manipulation cases.

3.3 Action and Control

Why think PPR is true? A potential source of motivation for the principle stems from the observation that performing an action requires some sort of control. As Frankfurt (1978) observes, “What is not merely pertinent but decisive [in determining whether the movements of a person’s body are actions of his] is to consider whether or not the movements as they occur are under the person’s guidance (p. 158). Similarly, Alvarez says that “for what someone does to be his action, for him to be its agent, the person must have a certain degree of control over it.” These remarks are doubtlessly correct. If a person has absolutely no control over the relevant event—a particular movement of his hand, for example—it is plausible to suppose that the event in question is not among the person’s actions. But in what, exactly, does the requisite guidance or control consist? According to Alvarez, it consists minimally in the person being “capable of refraining from doing what he does” (pp. 75-76).

If S A-s and it was within S’s power to refrain from A-ing, then S had control over whether or not he A-ed. Apparently, however, a person may control an event without having control over whether or not it occurs. Imagine a driving instructor who is happy to let his student steer the car right, but if the student shows any inclination of steering the car left or of continuing straight, the instructor is determined to take control of the car and steer it right. As it happens, the student steers the car to the right entirely of her own accord. By doing so, she guides the car in a certain direction, thus controlling the movement of the car, but given the instructor’s intention, it seems she lacks control over whether or not the car goes left or right. In this case, the control the student exercises over the movement of the car consists not in having control over whether the car goes right or left, but rather in her causing the car to go right.\footnote{The example is from Fischer and Ravizza 1998, p. 32. They use it to distinguish between two different kinds of control, and they go on to argue that just because a person lacks one kind—the kind that involves having access to alternative possibilities—it does not follow that the person lacks the kind of control...}
Examples like this illustrate an important point: there are different kinds of control, and a person may lack one kind of control while retaining other kinds. The observation that action requires some sort of control therefore does not automatically establish the further thesis that action requires a kind of control that involves having access to alternative possibilities. To establish this further thesis we would need additional reasons to think that possession of the relevant sort of control is or essentially involves the ability to refrain. However, no such reasons have been provided by proponents of PPR. So we can agree with Alvarez that if $S$ has no control over $A$, $A$ is not among $S$’s actions, but we needn’t follow her in assuming that the requisite control consists minimally in the ability or power to refrain. We might instead agree with Frankfurt that “The assertion that someone has performed an action entails that his movements occurred under his guidance, but not that he was able to keep himself from guiding his movements as he did” (1978, p.161).

Appreciating these points can also help rebut an objection to Frankfurt-style cases. The objection can be summarized as follows:

4. If Frankfurt-style cases are genuine counterexamples to the likes of PAP, then in the counterfactual sequence of events in which Black intervenes and forces Jones’s hand, it must be the case that Black causes Jones to perform the same action (or at least an action of the same type) that Jones performs in the actual sequence of events (premise).\(^8\)

5. In the counterfactual sequence Black does not cause Jones to perform an action; \textit{a fortiori}, then, Black does not cause Jones to perform the same action (or an action of the same type) that Jones performs in the actual sequence of events (premise).

6. Therefore, Frankfurt-style cases are not genuine counterexamples to PAP (from 4 and 5).

Suppose 4 is true. Even so, why think 5 is true? Alvarez insists that “in the counterfactual case Black must have total control over what happens,” and, consequently, that “Jones

\(^8\) See Alvarez 2009 pp. 65-67
cannot have any control over it,” in which case, she says, “what happens cannot count as an action of Jones’s, as something of which he is the agent—at most it might count as movements of his limbs or as events in his brain” (p. 77). But this sort of defense of 5 is incomplete at best, and it should now be clear why. It is true that in the counterfactual sequence Black has total control over whether the events in question occur and that Jones therefore lacks the kind of control over these events that involves having access to alternative possibilities, but as we’ve seen, it doesn’t automatically follow from this that Jones lacks all forms of control. In particular, it does not follow—at least not obviously and without further argument—that Jones lacks the control necessary for performing an action. The observation that performing an action requires some sort of control and that Jones lacks one kind of control therefore does not all by itself establish the truth of 5. In addition, it would need to be shown that the sort of control Jones lacks is the sort required for action, and this Alvarez has not done.⁹

### 3.4 Action and Inability

Are there, or could there be, actions from which agents are unable to refrain? If there are, perhaps the most natural place to begin searching for them is in cases of severe addiction or various neurological or psychological disorders. A minority of those suffering from Tourette’s syndrome experience coprolalia, uncontrollable outbursts of obscene or socially inappropriate remarks; Obsessive Compulsive Disorder (OCD) results in an array of compulsive behaviors; and dipsomaniacs are said to suffer from periodic, irresistible cravings for alcohol. Whether and to what extent these and other conditions actually render agents unable to refrain from performing certain actions is disputable. However, proponents of PPR must not only deny that such conditions ever actually render agents unable to avoid performing certain actions, they must also deny that it is even possible for an addiction or other condition to thus incapacitate a person. Both denials are striking, even after we have allowed for the fact that agents who suffer from various disorders and addictions may sometimes be able to refrain from the compulsive behavior associated with their condition.

⁹ For a different defense of 5, see Larvor 2010. I take up his argument later in this chapter.
Bob is a compulsive hand washer. On one particular occasion he acquired a very strong urge to wash his hands immediately, and circumstances were friendly to his satisfying this urge (e.g., a functioning faucet and soap were nearby). At the time he had no motivation to refrain from washing or to do anything incompatible with his washing then, and his OCD together with the circumstances at the time rendered him incapable of acquiring any such motivation. (This is compatible, of course, with his being capable then of acquiring the relevant motivation at some other time under different circumstances, a point to which I return shortly.) Unsurprisingly, Bob washed his hands, and his doing so is plausibly counted among his intentional actions.

Could Bob have refrained from washing his hands on this particular occasion? Given the details of the case, I suspect that many of us would be inclined to say he could not. (We can safely set aside the remote possibility that Bob might unintentionally refrain from washing.) Why? Because a necessary condition of his doing so—the presence of relevant motivational states—is absent, and, given the circumstances at the time, cannot be obtained. In order to intentionally do anything other than washing, Bob would need motivation to do otherwise. Since he lacks any such motivation and is presently incapable of acquiring it, it seems plausible to suppose that he is unable to intentionally do otherwise than wash his hands.

Suppose that, contrary to fact, Bob had managed to intentionally do otherwise than wash his hands on this occasion. What might explain this? No doubt a change in motivation would need to be part of the explanation, but apparently Bob is unable to affect any such change. That he wants to wash his hands immediately is presumably not up to him, and by hypothesis his OCD together with the circumstances at the time rendered him incapable of acquiring motivation to refrain from washing or to do anything incompatible with his washing then. So if, in this counterfactual scenario, Bob did somehow manage to intentionally do otherwise than wash, his doing so apparently cannot be explained by any change in his motivational states. But then what could explain his intentionally doing otherwise if not a change in motivation? In the absence of an explanation in terms of motivational change, and barring any special assumptions, it would seem that Bob’s intentionally doing otherwise, given his present motivational state, would simply be inexplicable.
Both Alvarez and Steward are aware of the challenges posed to their preferred theory of action by cases like this, though they have different ways of attempting to defuse it. Alvarez suggests that in cases of addiction and OCD “the truth seems to be that, for any particular occasion, the agent can avoid acting. The compulsion...lies not so much in each action but in a pattern of irrational...repetition” (p. 73). She goes on to note that “these conditions are often treated with cognitive-behavioral therapies that work by increasing awareness of the behaviour and its psychological origins, replacing the behaviour with other activities, etc., all of which are aimed at removing the habit of repetition” (p. 74). But according to Alvarez, such therapies “could not even begin to be deployed if the actions that form the pattern were unavoidable” in the sense at issue in PAP (p. 74).

As it happens, Bob recently sought therapy for his OCD and, consequently, no longer engages in compulsive hand washing. His treatment involved the sort of cognitive-behavioral techniques Alvarez mentions. During sessions with his therapist Bob would often be placed in circumstances of the sort that triggered severe anxiety within him. At such times he would typically have resorted to repeated hand washing, but during these sessions he lacked the opportunity to engage in this behavior. Instead, his therapist helped him attend to the thoughts and feelings evoked by the circumstances in question and taught him various techniques that enabled him to cope with these thoughts and feelings without engaging in compulsive behaviors. Although this type of therapy was mentally excruciating for Bob at first, over time he eventually learned how to deal with the anxiety aroused by these circumstances without resorting to repeated hand washing.

Does the success of Bob’s treatment require or in any way suggest that he was able to refrain from compulsive hand washing prior to receiving treatment? In particular, does it imply that, in the episode related earlier, Bob was able to intentionally refrain from that particular act of hand washing? To insist that it does ignores the possibility that behavior that is avoidable for a person at one time and in one set of circumstances might be unavoidable for that same person at some other time in a different set of circumstances. The fact that, as a result of extended counseling, a compulsive hand washer like Bob can learn how to cope with his anxieties without resorting to problematic behaviors is consistent with those same behaviors being unavoidable for him prior to his
receiving treatment. Presumably, Bob couldn’t read prior to learning how to do so. Perhaps he also couldn’t resist washing his hands under certain circumstances prior to his therapist teaching him how to cope with the anxiety prompted by those circumstances.

Steward suggests a different approach to these matters. She grants that, owing to an addiction or other underlying condition, a person may be unable to avoid engaging in a certain type of behavior within a given time frame. However, according to her, “That the occurrence of a behavior-type has absolutely no chance of not occurring within a certain time-frame under certain conditions, does not imply that the token action which constitutes its actual occurrence” was unavoidable (2008, p. 142). To illustrate her point, Steward considers an alcoholic who has easy access to a bottle of whisky. She says that, because of his addiction, the alcoholic may not be able to resist drinking the whisky at some time or other. Whisky drinking behavior is thus bound to occur sooner or later. However, Steward contends that the details about how the alcoholic actually goes about drinking the whisky—for example, where he drinks it, how he drinks it, the precise moment at which he drinks, etc.—will be up to him, in which case, she claims, the token act of whisky drinking that the alcoholic eventually performs was not unavoidable, “but rather a genuine exercise of the two-way power of agency, as all true actions must be” (p. 142). A similar account might be given regarding compulsive behavior of the sort Bob used to engage in.

Steward’s handling of addiction and compulsion fares better than Alvarez’s, but problems remain. Steward’s view seems to require that actions be extremely fragile events in the sense that a slight change in how or when an action is performed results in a different token action. Ordinarily, however, we do not treat most other events in this way. As David Lewis once pointed out, “we are usually quite happy to say that an event might have been slightly delayed, and that it might have differed somewhat in this or that one of its contingent aspects” (2000, p. 185). How much change an event can undergo and yet remain the same event is of course a tricky question. However, if we allow, as we ordinarily do, that a particular event could have occurred in a slightly different way or could have begun at a slightly different time than it actually did and yet remain the same event, the fact (if it is indeed a fact) that an addict or someone suffering from OCD or some other disorder was able to refrain from A-ing-in-a-particular-way will not itself
support the conclusion that he could have refrained from A-ing. Thus, if Steward’s defense of PPR is to succeed, she must also defend the view that events, or a certain subclass of events like actions, are fragile entities. More importantly, however, even if the view that actions are fragile is defensible, it is doubtful whether Steward can exploit this fact in defense of PPR in all problematic cases arising from addiction or compulsion.

Addicts and those suffering from various neurological and psychological disorders may, as a result of their condition, be unable to avoid performing certain actions. However, even psychologically healthy agents arguably find themselves in circumstances in which they are unable to refrain from performing some action. In fact, in certain cases it may be argued that it is hallmark of psychological health that a person cannot avoid acting in a particular way. It is said, not implausibly, that the fully virtuous agent cannot avoid doing the virtuous thing, at least not intentionally. Perhaps there are in fact no fully virtuous agents in this sense, but certainly such agents seem possible. Moreover, there are arguably some things that even those of us who fall just shy of complete virtue cannot bring ourselves to refrain from doing. To see this, it will be helpful to first briefly examine a special class of actions.

Some actions are simply unthinkable for a person. As I understand unthinkable actions, if A-ing is unthinkable for S at t, then, everything else being equal, at t, S is unable to intentionally A. It certainly seems as if there are some actions that are unthinkable in this sense. Daniel once claimed to be unable to torture an innocent victim in return for a small sum. I too believe myself incapable of committing such atrocities, given my present psychology. Neither Daniel nor I presently have any motivation to torture innocents; that is, we do not presently want (either intrinsically or extrinsically) to perform any such action. Moreover, we may suppose, plausibly, that our revulsion (on a number of levels) to the idea of performing such actions prevents us from acquiring the pertinent motivation. Now, if S has no motivation at t to A, and if S is incapable, at t, of acquiring motivation to A, then arguably S is unable to intentionally A just then. Supposing this is correct, it follows that both Daniel and I are unable just now to

11 With one exception, “t” is a period of time longer than a moment. The exception is in the case of momentary actions such as decisions.
12 See Dennett 1984, p. 133.
intentionally torture an innocent victim for small amounts of money. That is just something that, given our circumstances, including our present psychology, we could not bring ourselves to do. This is a rather vivid illustration, to be sure; no doubt more homely examples could be adduced.¹³

A reminder about ability claims is in order. The sense of ability at issue in the present discussion is something stronger than a general ability to A, where this is understood as a disposition, capacity, skill or competency to A. That A-ing is unthinkable for S at t is consistent with S retaining, at t, the general capacities needed to A right then. But if A-ing really is unthinkable for S at t in the sense presently at issue, then S is unable to exercise the relevant capacities at t in an attempt to A. Daniel no doubt retains certain general physical abilities of the sort that would be employed in torture, and so do I. Presumably, however, what Daniel is claiming when he claims to be unable to torture an innocent victim in return for a small sum is that he could not right now bring himself to exercise these abilities for that purpose. This is certainly what I am claiming when I claim to be unable to torture an innocent victim for a small sum.

If some actions are unthinkable in the sense presently under discussion, some omissions are arguably unthinkable in this sense, too. Unthinkable omissions can be understood in much the same way as unthinkable actions. If intentionally omitting to A is unthinkable for S at t, then, everything else being equal, at t, S is unable to intentionally refrain from A-ing. Are there, or could there be, unthinkable omissions in this sense?

Consider Esther, whose child is in agonizing pain and is begging for relief. Suppose that Esther is the only one who can relieve the child’s suffering and that the only way she can do so is by pressing a certain button at t, thereby intravenously delivering a powerful pain medication to the child. Esther, being a good and loving parent, wants to press the button at t in order to relieve her child’s suffering, and circumstances are friendly to her pressing it (e.g., she is capable of pressing the button, there is no external barrier preventing her from pressing it, she isn’t ignorant of the button’s location, etc.). Moreover, she currently has no motivation to refrain from pressing the button or to do anything incompatible with her pressing it at the relevant time, and, because she loves her child so much, she is psychologically incapable of acquiring such motivation just now.

Refraining from pressing the button at $t$ and thereby allowing her child to continue to suffer in agony is, for Esther, in those circumstances, simply unthinkable. Predictably, Esther presses the button at $t$. Was she able to refrain from pressing it? Given the details of the case, I suspect that many of us would be inclined to say she was not. (Again, we may safely set aside the remote possibility that Esther might unintentionally refrain from pressing the button at $t$.) And yet, pressing the button at $t$ is plausibly counted among Esther’s intentional actions.

### 3.5 Whose Action?

According to Brendan Larvor (2010), Frankfurt-style cases fail to refute PAP because in the counterfactual scenario in which Black intervenes and forces Jones’s hand, it is Black, and not Jones, who kills Smith. Thus, Larvor concludes, Jones can avoid killing Smith after all by unwittingly making Black intervene. And if Jones can avoid killing Smith, then clearly we do not have a counterexample to PAP.

Suppose we grant that, in the counterfactual scenario, Black kills Smith. But notice that Black killing Smith is compatible with Jones killing Smith, as well. To defuse Frankfurt’s attempted counterexample, then, it is not enough to point out that Black kills Smith in the counterfactual scenario. In addition, one must show that Jones does not also kill Smith, and this, I will argue, is something Larvor has not shown. Moreover, there are actions other than killing Smith that Jones apparently cannot avoid performing but for which he can perhaps be morally responsible.

Why deny that, in the counterfactual scenario, Jones kills Smith? Larvor begins by observing that we can revise our decisions. I might decide this morning to go to the store later this evening only to change my mind in the afternoon. But how is this relevant to Frankfurt-style cases? Suppose Jones is disinclined to kill Smith. Then Black will intervene and make Jones decide to kill Smith. Now, what if Jones’s own inclinations lead him to revise this decision? In that case, Black will have to intervene again, making Jones re-decide to kill Smith. If this flip-flop continues right up to the time of the shooting, Black will be forced to intervene at the very last moment. But then, Larvor says,
“there is no difference between making Jones decide to pull the trigger and making Jones pull the trigger...Jones is merely Black’s unwilling instrument” (2010, p. 507).\(^{14}\)

These statements are puzzling. A decision by \(S\) to \(A\) is commonly and plausibly taken to be an intentional action consisting in the formation by \(S\) of an intention to \(A\).\(^{15}\) Given this widely held view of decisions, if, as Lavor apparently concedes, Black can send a signal to Jones’s brain and thereby make Jones decide to shoot Smith, there is indeed an important difference between making Jones decide to pull the trigger and simply using his body as an instrument to pull it. Were Black to make Jones decide to pull the trigger, he would make Jones perform an action, whereas if Black simply uses Jones’s body as an instrument to pull the trigger while bypassing Jones’s will, he does not make Jones perform an action.

Lavor assumes that if Black had intervened because “Jones’s own inclination is against doing the deed,” that Jones would have found “his limbs going through the motions,” in which case, Lavor insists, “the deed ceases to be his” (2010, pp. 507-508). Now, if Black’s intervention has to work by bypassing Jones’s will (i.e., his beliefs, desires, inclinations and intentions), then, to be sure, Jones is an unwilling instrument and what happens in the counterfactual scenario is not his action. But it is generally—and, I think, correctly—assumed that Black could intervene in a way that does not bypass Jones’s will. If Black causes Jones’s limbs to move indirectly by first causing Jones to acquire pertinent beliefs, desires, inclinations and intentions that in turn deterministically and non-deviantly result in the relevant bodily movement (say, the pulling of a trigger), then, contrary to what Lavor claims, it is quite plausible to suppose that what happens can be among Jones’s actions.

Lavor focuses on overt actions, such as killing Smith. However, there are other actions that Jones performs in the scenario, as well. For example, the decision to kill Smith is a candidate for an action that Jones cannot avoid performing but for which he can perhaps be morally responsible. Apparently, Jones cannot avoid deciding to kill Smith any more than he can avoid killing Smith, for he either decides on his own to kill Smith or as a result of Black’s intervention. But since he decides on his own, many would

\(^{14}\) Question: why couldn’t Black simply prevent Jones from revising his decision? Lavor doesn’t say.

\(^{15}\) For a helpful discussion of deciding, see Mele 2000.
say that he can be responsible for this decision nonetheless, and nothing Lavor says challenges this line of argument.

Perhaps Larvor will insist that whatever Black causes in the counterfactual scenario cannot be an action of Jones’s, whether it is an overt bodily movement like pulling a trigger or a mental event such as the formation of an intention to pull the trigger. If Lavor could establish this claim, he will indeed have succeeded in defusing Frankfurt’s counterexample. However, Larvor offers no argument for the claim that Black cannot deterministically cause Jones to perform an action.

Is there any reason to doubt that Black could deterministically cause Jones to perform an action in such a way that Jones could not avoid performing it? If PPA were true, then we would have just such a reason. However, as I have argued, there is little, if any, reason to think PPA is true and good reason to think it is false. And, in the absence of something like PPA, I see little reason to accept the claim that Black cannot bring it about in the counterfactual scenario that Jones performs an action.

### 3.6 Two Implications

PPR says that an event $A$ is an action of $S$’s only if $S$ could have refrained from $A$-ing. I’ve argued that this principle is both insufficiently motivated and subject to credible doubt in a range of cases. I conclude that performing an action does not require the ability to refrain from performing it. My conclusion has two important implications. First, each of Alvarez’s, Steward’s, and Lavor’s recent attempts to defend PAP are unsuccessful. If PPR is false, then the argument for PAP suggested by Steward (discussed in section 2.2) is unsound, and whatever other problems there may be with attempts to falsify PAP using Frankfurt-style cases, once we jettison PPR, there is little, if any, reason to think that a “successful” Frankfurt-style counterexample to PAP must presuppose a conceptual impossibility. Second, that at least some actions could be unavoidable lends at least some support to the idea that the notion of an action is, as Fischer puts it, an actual sequence notion. According to actual sequence theorists about action such as Fischer and myself, what makes an event an action has primarily to do with the nature of the event in question.
and what happens in the actual sequence of positive events leading up to that event, not on whether the agent had access to alternative possibilities with respect to that event.\footnote{Some might object that I am not entitled to this second conclusion until I falsify views according to which what makes an event an action is something intrinsic to that event (e.g., Ginet 1990). Such views do not count as actual sequence theories, since an event’s being an action, on these views, will not depend on its relation to past events. I think this is a fair criticism. Still, the arguments of this paper constitute a partial defense of actual sequence theories insofar as they succeed in showing that an agent needn’t have access to alternative possibilities in order to perform an action. A comprehensive defense of an actual sequence theory would need to deal with this third type of theory as well.}
CHAPTER 4

BLAMEWORTHINESS WITHOUT WRONGDOING

It is widely acknowledged that, owing to the possibility of genuine excuses, a person can do something morally wrong without being blameworthy. Kit accidentally kills a kitten; he backs his car out of his garage, unaware that the kitten is taking a snooze behind his back tire. Kit has done something he shouldn’t have—he oughtn’t to have killed the kitten—but since he was ignorant of the kitten’s whereabouts at the time, he is not blameworthy for what he did (assuming, of course, that he is not to blame for his ignorance).\(^1\) Wrongdoing therefore does not entail blameworthiness.\(^2\) However, many believe that the reverse entailment does hold—that blameworthiness for actions (and omissions) entails wrongdoing. What are we to make of this claim?

At first blush it seems quite plausible indeed. It captures the intuitive but somewhat inchoate thought many of us have that blaming a person for his behavior is appropriate only insofar as there is something morally untoward about what he did. It’s difficult to imagine how someone could be blameworthy for doing something that was in no way morally objectionable. However, matters are a bit more complicated than they appear, for there are at least three distinct claims one might be making when one says that blameworthiness entails wrongdoing, only one of which is true.

Philosophers sometimes distinguish between objective and subjective wrongdoing. Objective wrongdoing occurs when, and only when, a person either does something he has a moral obligation not to do, all things considered, or fails to do something he has a moral obligation to do, all things considered. Whereas objective wrongdoing occurs when agents do things they shouldn’t, subjective wrongdoing occurs when agents do things they believe they shouldn’t. Thus, subjective wrongdoing occurs when, and only when, a person either does something he believes he has a moral

---

1 The example is from Fischer and Ravizza 1998, p. 13.
2 For a defense of these claims, see Austin 1961 and Zimmerman 2004.
obligation not to do, all things considered, or fails to do something he believes he has a moral obligation to do, all things considered.

Objective and subjective wrongdoing typically occur together; an agent who believes he is doing something he shouldn’t is often correct. Nevertheless, we might have one without the other. When Kit ran over the kitten, he did something he shouldn’t have—he killed the kitten—but presumably he didn’t believe at the time that he was doing something wrong. Thus we have a case of objective wrongdoing without subjective wrongdoing. Might there also be cases of subjective wrongdoing without objective wrongdoing? I think so, but we’ll come to that later.

Those who insist that blameworthiness requires wrongdoing typically have objective wrongdoing in mind. Call this the Objective View of blameworthiness. The Objective View, I will argue, is false; it is possible for someone to be blameworthy for what he did even if he didn’t do anything objectively wrong. Having recognized this, several philosophers have insisted that, although blameworthiness may not entail objective wrongdoing, it does entail subjective wrongdoing. Call this the Subjective View of blameworthiness. The Subjective View is false, as well; it is possible for someone to be blameworthy for what he did even if he didn’t do anything subjectively wrong. What doesn’t seem possible, however, is for someone to be blameworthy for doing something that was neither objectively wrong nor subjectively wrong. The most plausible view concerning the relationship between blameworthiness and wrongdoing is therefore disjunctive: if a person is blameworthy for what he did, then his action was either objectively or subjectively wrong. My aim in what follows is to defend these assertions.

4.1 The Objective View

Let’s start with the Objective View:

**Objective View of Blameworthiness**: Necessarily, for any action or omission \( \phi \), if an agent \( S \) is blameworthy for \( \phi \)-ing, then it was objectively wrong for \( S \) to \( \phi \).

---

3 Here I follow Haji 1998.
As I said, this view is quite popular, and has been put to a number of very different, and sometimes opposing, philosophical uses. For instance, it figures prominently in the following argument for the incompatibility of determinism and blameworthiness:

1. Necessarily, if \( S \phi \)-ed and determinism is true, then \( S \) couldn’t have avoided \( \phi \)-ing (assumption: incompatibility of determinism and the ability to do otherwise).
2. Necessarily, if \( S \) couldn’t have avoided \( \phi \)-ing, then \( S \) did not have an all things considered obligation not to \( \phi \) (from the principle that ‘ought’ implies ‘can’).
3. Necessarily, if \( S \) did not have an all things considered obligation not to \( \phi \), then it was not objectively wrong for \( S \) to \( \phi \) (true by definition).
4. Necessarily, if it was not objectively wrong for \( S \) to \( \phi \), then \( S \) is not blameworthy for \( \phi \)-ing (from the Objective View).
5. Therefore, necessarily, if \( S \phi \)-ed and determinism is true, then \( S \) is not blameworthy for \( \phi \)-ing (from 1-4).

It has also been used to defend the Principle of Alternative Possibilities,

\textbf{PAP}: \( S \) is blameworthy for performing an action \( A \) only if \( S \) could have avoided \( A \)-ing.

Here’s the argument:

1. \( S \) is blameworthy for \( A \)-ing only if it was objectively wrong for \( S \) to \( A \) (from the Objective View).
2. It was objectively wrong for \( S \) to \( A \) only if \( S \) had an all things considered obligation not to \( A \) (true by definition).
3. \( S \) had an all things considered obligation not to \( A \) only if \( S \) could have avoided \( A \)-ing (from the principle that ‘ought’ implies ‘can’).


\footnote{An argument along these lines is suggested by van Inwagen 1983, especially pp. 161-162. However, I’m not sure whether van Inwagen would endorse the argument presented here.}
4. Therefore, PAP is true (from 1-3).  

Finally, it figures in an argument of John Martin Fischer’s against the venerable deontic maxim that ‘ought’ implies ‘can’,

**OIC**: Necessarily, if $S$ has an all things considered moral obligation to $\phi$, then $S$ can (is able or has it within his power to) $\phi$.

Fischer’s argument begins with a case in which Sally opts not to raise her hand even though she knows that doing so would alert a lifeguard to a drowning child. Unbeknownst to Sally, however, she is temporarily paralyzed and thus cannot raise her hand. In Fischer’s opinion,

[Sally] is morally responsible (and blameworthy) for not raising her hand, even though she could not have raised her hand. Further, since she is blameworthy for not raising her hand, I would claim that she acted wrongly in failing to raise her hand, and thus that she ought to have raised it. But she could not have raised it. Thus, ought does not imply can. (Fischer 2006, pp. 24-25)

Fischer’s argument here seems to be this:

1. Sally is blameworthy for not raising her hand (this is supposed to be intuitive).
2. Therefore, it was objectively wrong for her not to raise it (from 1 and the Objective View).
3. If it was objectively wrong for Sally not to raise her hand, then she had an all things considered obligation to raise her hand (true by definition).
4. Therefore, Sally had an all things considered obligation to raise her hand (from 1-3).
5. Sally couldn’t raise her hand (true by hypothesis).

---

6 Versions of this argument have been defended by Copp 1997 and 2003, and Speak 2005. The argument was first suggested by Widerker 1991.
6. Therefore, OIC is false (from 4 and 5).

Notice that if this argument is sound, then the previous two are not, since both rely not only on the Objective View but on OIC, as well.

Each of the preceding arguments relies on premises other than the Objective View that can be questioned. However, the Objective View is central to all three. So if, as I will argue, it is false, then all three arguments can be undermined in one fell swoop. Also, given the plausible and widely held view that objective wrongdoing does not suffice for blameworthiness, owing to the possibility of excuses, if the Objective View is false, it follows that blameworthiness and objective wrongdoing are conceptually independent of one another. That in and of itself would be an interesting and important conclusion.

Given the popularity of the Objective View, it might be worthwhile to pause briefly and ask why so many philosophers are attracted to it. Part of the answer, I believe, is that in thinking about blameworthiness we often fail to consider a sufficiently diverse range of cases. In paradigmatic cases of blameworthiness the action or omission for which the agent is blameworthy is often objectively wrong. Blameworthiness and objective wrongdoing thus typically go hand in hand. So it might be tempting to generalize and say that all cases of blameworthiness must involve such wrongdoing. The temptation should be resisted, however. If we focus only on typical cases of blameworthiness, cases in which the agent is blameworthy for doing something he was morally obligated not to do, the Objective View will seem quite attractive indeed. But we should not ignore the possibility that there are atypical cases in which the agent is blameworthy for what he did or failed to do despite not having done anything objectively wrong. The fact that blameworthiness and objective wrongdoing typically go together therefore does little to support the Objective View. At best, it shows that there may be an initial presumption in favor of the view, one that might be easily overturned upon consideration of a wider range of cases.

Perhaps another reason some have been attracted to the Objective View is that they have run together act and agent evaluation. Bad behavior is often grounds for a negative assessment of the agent, but not always; act and agent evaluation can be pried apart. This is illustrated by the possibility of excuses. Remember Kit’s mishap with the
kitten. Recall that Kit is not blameworthy for killing the kitten, but just because a negative assessment of Kit is not warranted in this instance, it does not follow that what he did was morally permissible. The fact that he didn’t know the kitten was behind his tire doesn’t make backing over it morally permissible. Given that objectively wrong behavior does not necessarily warrant blaming the agent for what he did, I suggest that we should at least be open to the possibility that act and agent evaluation might be pried apart in the other direction. That is, we should be open to the possibility that blaming an agent for his behavior might be warranted even if the person didn’t do anything objectively wrong.

Are there any further considerations in support of the Objective View? One argument that suggests itself stems from further reflection on cases of non-culpable ignorance. Why isn’t Kit blameworthy for killing the kitten? An initially plausible answer is that he didn’t know, and, let us assume, couldn’t have been expected to know, that the kitten was there. Reflection on cases like this therefore might be taken to suggest that, barring culpable ignorance, a person is blameworthy for what he did only if he knew at the time that he was doing something objectively wrong. Since knowledge entails truth, it follows that a person is blameworthy for what he did only if he did in fact do something objectively wrong.

This argument is unsound, however. The fact that ignorance of the pertinent features of one’s actions often excuses should not lead us to believe that, cases of culpable ignorance aside, knowing that you are doing something objectively wrong is required for being blameworthy for what you did. There are various ways that S might fail to know that p; it might be false that p, S might fail to believe that p or, which is the relevant point for present purposes, S might not be justified in believing that p. Suppose that S has an all things considered obligation not to A, correctly believes that A-ing is objectively wrong, but for whatever reason lacks justification for this belief. In that case, S does not know that A-ing is objectively wrong, but S might be blameworthy for A-ing nonetheless.

Consider the following example.
KILL BILL: Beatrix believes (correctly, let us assume) that she has an all things considered moral obligation not to kill her ex-boyfriend, Bill, and therefore that it would be objectively wrong to kill him. But, let us suppose, she is not justified in believing this. She has some evidence that killing Bill would be permissible in this case (perhaps she has evidence that he is about to kill her daughter and that the only way to prevent him from doing so is to kill him), but she doesn’t bother looking into the matter further because, although she is an otherwise normal individual, she has a strange belief that it is always wrong to kill people whose names begin with the letter ‘B’. In the end, Beatrix freely kills Bill, solely to inherit his fortune, despite believing that it was objectively wrong to do so.

Given the strange basis of Beatrix’s belief that it is objectively wrong to kill Bill, it would be a mistake to say that this belief of hers is epistemically justified. And, because Beatrix is not epistemically justified in believing that killing Bill is objectively wrong, it would also be a mistake to say that she knew that it is objectively wrong to kill him. Assuming she is not culpably ignorant, the claim that (cases of culpable ignorance aside) a person is blameworthy for what he did only if he knew that he was doing something objectively wrong yields the implausible result that Beatrix is not blameworthy for killing Bill.

Being the appropriate object of blaming attitudes, such as resentment, anger, indignation and guilt, is indicative of blameworthiness, and evidently Beatrix is the appropriate object of these attitudes. It would be appropriate for Bill’s loved ones to resent Beatrix for killing Bill. Likewise, it would be appropriate for us to be indignant toward Beatrix for what she did, and for Beatrix to feel guilt for her actions. It seems, then, that Beatrix is indeed blameworthy for killing Bill, even though she did not know at the time of action that killing him was objectively wrong. Knowledge of objectively wrongdoing is therefore not required for blameworthiness, even in cases in which culpable ignorance is not an issue. The most obvious defense of the Objective View is therefore unsuccessful.

One final reason some people may be attracted to the Objective View is because they see it as the only way to capture the intuitive thought mentioned earlier that blaming a person for his conduct is only appropriate if there is something morally untoward about
what he did. This is not a very good reason to accept the View, however, for it assumes, without argument, that neither the Subjective View nor the Disjunctive View has the resources to account for the relevant intuition. I return to this issue a bit later.

4.2 A Minority Report

Despite the popularity of the Objective View, there are a handful of philosophers who reject it.\(^7\) According to Ishtiyaque Haji, for instance, people can be blameworthy for performing suberogatory actions. An action is suberogatory provided that it is not objectively wrong but is nevertheless morally bad in some respect.\(^8\) If there are such things as suberogatory actions, and if, as Haji claims, people can be blameworthy for performing them, then contrary to what the Objective View implies, a person can indeed be blameworthy for what he did even if he didn’t do anything objectively wrong.\(^9\)

Haji recognizes that “Resistance to this line of argument…is to be expected from those skeptical of the possibility of suberogation,” but he contends that such skepticism is not warranted (2002, p. 173). I am not convinced. Despite the efforts of Haji and others to make the idea of suberogation more palatable, it remains suspect. I will not defend this claim here, however, for even if suberogatory actions are possible, Haji’s appeal to them in the present context strikes me as a bit of a red herring. Let me explain.

Grant for the sake of argument that suberogatory actions are possible. Why does Haji think that someone who performs such an action can be blameworthy for what he did? Commenting on Julia Driver’s (1992) example of a man who, being first in line, takes a more convenient seat on a train and thereby deliberately prevents two lovers from sitting together when he could have easily taken a less convenient seat, Haji says that if we assume the man “believes he is doing wrong by preventing the lovers from sitting together despite the availability of another seat,” then “may be blameworthy for not giving up his seat even though, let’s assume, his not giving up his seat is suberogatory and hence morally optional” (2002, p. 171). For the sake of argument, suppose Haji is right about all this. Even so, the fact that the man’s action is suberogatory doesn’t seem to

---

\(^8\) Cf. Driver 1992.
play a crucial role in the judgment that he is blameworthy. Rather, the reason Haji thinks the man is blameworthy seems to be that, by preventing the lovers from sitting together, the man deliberately did something he mistakenly believed to be objectively wrong. The fact (if it is indeed a fact) that the man did something suberogatory is only relevant insofar as it shows that the man’s belief that his action was objectively wrong was in fact mistaken. But whether he was mistaken because the action was indeed suberogatory, as Haji claims, or because it was merely permissible but not suberogatory seems to be entirely beside the point. What is essential to Haji’s argument, it seems, is that (not why) the man was mistaken in thinking he was doing something objectively wrong.

Haji provides a similar analysis about a case in which “Joy is aware that others are waiting for her table in the crowded café, mistakenly believes she is doing wrong by lingering on, intentionally lingers on in light of the belief that she is doing wrong, but really suberogates by lingering on” (2002, p. 39; Haji’s emphasis). He insists that Joy is blameworthy for lingering on, and again his reason seems to be that she deliberately did something she mistakenly believed to be wrong. But again, it doesn’t seem to matter whether she was mistaken because lingering on was in fact suberogatory, as Haji thinks, or because it was merely permissible but not suberogatory. What matters is that she was in fact mistaken in thinking that she was doing something objectively wrong but is blameworthy for what she did nonetheless.

So in the end it’s not clear that Haji’s appeal to suberogation is really doing much heavy lifting in his argument against the Objective View. Rather, the thought seems to be that if someone deliberately does something in light of a belief that he is thereby doing objective moral wrong, that person may be blameworthy for what he did, even if his belief that the action in question is objectively wrong is mistaken. Insofar as this is the main thrust of Haji’s argument, it is similar to an argument of Michael J. Zimmerman’s. 10 According to Zimmerman, it is possible for someone to be blameworthy for what he did even if he didn’t do anything objectively wrong, because “acting freely in the belief that one is doing objective moral wrong is sufficient for being morally blameworthy for one’s behavior,” but not for actually doing something objectively wrong (1997a, p. 234).

---

10 Brandt 1958, pp. 38-39 also seems to endorse this argument.
Zimmerman’s argument relies on two assumptions. The first is that freely performing an action in light of a belief that the action is objectively wrong is not sufficient for objective wrongdoing. The second is that freely doing something one believes to be objectively wrong is sufficient for being blameworthy for one’s behavior. I have no qualms with the first assumption. It is the second with which I take issue.

Consider the following story, which is inspired by an episode in Corrie Ten Boom’s book *The Hiding Place*.

**HIDING PLACE:** It’s 1943, and Corrie is hiding Jewish refugees in her home in Haarlem, Holland. One day a group of Nazi officers come knocking and ask Corrie if there are any Jews in the house. Now, Corrie takes the Ten Commandments very seriously, and, like Kant, believes that lying is always objectively wrong, no matter what. She therefore believes that it would be objectively wrong for her to lie to the officers. She also believes that it would be objectively wrong to tell them the truth, however, since by doing so she would effectively be handing over the refugees to be tortured and killed. Remaining silent is no good either, for then the officers will get suspicious, raid the house, find the people she is hiding and whisk them all, herself and her family included, off to concentration camps where they would all very likely perish. In short, no matter what Corrie does, she will do something she believes to be objectively wrong. In the end, although it pains her, morally, to do it, Corrie freely lies to the Nazis, despite believing that she is morally obliged not to do so, solely to prevent herself, her family and those she is hiding from being taken away and killed.

Two features of this case require emphasis. First, I have assumed that Corrie acted freely. This assumption is plausible. In the end, she decided to lie, but (let us suppose), she could have decided and done otherwise—e.g., she could have decided instead to simply remain silent, and she could have acted accordingly. Moreover, no one forced her to lie, nor was her decision and subsequent action the result of an irresistible desire. We can even suppose that Corrie was the agent-cause of her actions, if you like. Given these

---

11 See Zimmerman 1997a, pp. 231-233 for a defense of this claim.
assumptions, I think it is safe to say that Corrie acted freely. The second thing to notice is that Corrie believed (mistakenly, I think) that lying to the Nazis was morally wrong.

The crucial question is whether Corrie is blameworthy for what she did. Because she freely lied despite believing that it was objectively wrong to do so, Zimmerman’s second assumption—i.e., that freely doing something in light of a belief that it is wrong is sufficient for blameworthiness for one’s behavior—implies that she is. However, I find this implausible. When a person is blameworthy for her conduct, the actions in question reflect ill on her and serve as a basis for a certain kind of negative moral judgment or appraisal of the person, at least with respect to that particular episode of her life.\textsuperscript{12} As Zimmerman puts it elsewhere, it’s as if there is debit or discredit on the person’s “moral ledger” in light of that episode in her life.\textsuperscript{13} However, Corrie’s action in this case does not reflect ill on her, nor does she merit any kind of negative moral appraisal on the basis of what she did. No bad marks should be entered on her moral report card. Although she did indeed freely do something she regarded as objectively wrong, she arguably did not do so in a way that incurs moral blame.

Someone who is blameworthy for what she did is, in principle at least, the appropriate object of the blaming attitudes. However, Corrie is arguably not the appropriate object of these attitudes. We have no cause to be indignant toward Corrie for what she did, nor should Corrie feel guilty about her behavior. This isn’t to say that it would be inappropriate for Corrie to feel bad about what happened. Perhaps she might reasonably regret, or otherwise find unsavory, the fact that she has to lie in order to save her friends, but given the details of her situation, she surely has no reason to reprimand, censure or blame herself for what she did.

For these reasons, it seems to me that Corrie is not blameworthy for lying to the Nazis. Since Zimmerman’s second assumption implies that she is blameworthy for doing so, I think we should reject that assumption; freely doing something in the belief that one is thereby doing objective moral wrong is not sufficient for being blameworthy for one’s conduct. Zimmerman has therefore failed to provide us with sufficient conditions for blameworthiness, and so, \textit{a fortiori}, has failed to provide us with sufficient conditions for

\textsuperscript{12} Cf. Zimmerman 1997a, pp. 229-230.
blameworthiness that are not also sufficient for objective wrongdoing. Zimmerman’s argument against the Objective View is therefore unsuccessful, and insofar as Haji’s argument is similar to Zimmerman’s, it too fails for much the same reason.

Although I find Haji’s and Zimmerman’s arguments unsatisfactory, I think they were on the right track. Like Haji and Zimmerman, I think people can sometimes be blameworthy for their subjectively wrong actions even if those actions are not objectively wrong. My aim in the following section will be to defend this thesis.

4.3 Joining the Minority Report

Although I am not persuaded by the arguments of Haji and Zimmerman, I too wish to join the minority report on the Objective View. Consider a variation on an earlier example.

**Kill Bill 2**: Beatrix believes she has an all things considered moral obligation not to kill her ex-boyfriend, Bill, and therefore that it would be objectively wrong to kill him. She also believes that she could avoid wrongdoing by simply not killing him. She freely kills him anyway, however, solely to inherit his fortune. Unbeknownst to Beatrix, Bill was about to torture and kill her daughter, and the only way she could have prevented him from doing so was to kill him.

I submit that this case is a counterexample to the Objective View. Beatrix is blameworthy for killing Bill, but it was not objectively wrong for her to do so.

Start with the claim that Beatrix is blameworthy for killing Bill. Why think that? Well, for starters, we might begin by pointing out that Beatrix is morally criticizable for killing Bill, insofar as she thereby freely did something she regarded as objectively wrong despite believing that she could avoid moral wrongdoing by not killing him. Her behavior in this case reflects ill on her; it is grounds for a debit in her moral ledger, as it were. If, as Zimmerman and others think, this suffices for blameworthiness, then Beatrix is indeed blameworthy for killing Bill.
Beatrix is also evidently the appropriate object of the blaming attitudes for having killed Bill. It would, it seems to me, be appropriate for us to be angry with and feel indignation toward Beatrix for killing him, insofar as she did so freely despite being aware of what she took to be, and what normally would have been, good and sufficient reasons not to kill him and despite believing that she could have avoided wrongdoing entirely by not doing what she did. For similar reasons, it would be appropriate for Beatrix to feel guilty for killing Bill. If being the appropriate target of the blaming attitudes for having A-ed is sufficient for blameworthiness for A-ing, then Beatrix is blameworthy for killing Bill.

Contrary to what the Objective View implies, however, it was arguably not objectively wrong for Beatrix to kill Bill. Central to my defense of this claim is the Principle of Self-Defense:

**PSD:** It is permissible to kill someone who, without good reason, is trying to kill you or those you have a duty to protect, (e.g., your children) if doing so is the only way to save your own life or the lives of those you are responsible for protecting, and if you can do so without doing something else that you are morally required not to do (e.g., killing an innocent bystander).

This principle, or something very much like it, is both intuitively plausible and widely accepted, and given several further assumptions, it implies that it was morally permissible for Beatrix to kill Bill.

Not everyone accepts PSD, however. Michael Otsuka (1994) argues that it is morally impermissible to kill an *innocent aggressor*—i.e., someone who endangers your life because she is trying to kill you, but who is not blameworthy for her behavior. Consider an example of Otsuka’s in which “someone...is pursuing you with a meat cleaver because she is moved by an uncharacteristic and overwhelming rage that has been induced by a powerful mind altering drug that someone has [secretly] slipped into her morning coffee” (1994, p. 75). Because the woman in this example is trying, without good reason, to kill you, PSD implies that it is permissible for you to kill her, assuming that doing so is the only way to prevent her from killing you and that you can do so
without violating any further moral requirements to which you might be subject at the
time. However, Otsuka contends that, because the woman is an innocent aggressor, it
would be impermissible to kill her even to save your own life. If he is right, then PSD is
false.

I’m not convinced by Otsuka’s argument, but never mind that. Even if he is right,
there is a version of PSD that is consistent with his position and that is also suitable to my
purposes here. Let us define an unjust aggressor as someone who is trying, without good
reason, to kill you or those, such as your children, who you have a duty to protect, and
who is blameworthy for doing so. Here is a version of PSD restricted to unjust
aggressors:

\[
\text{PSD*}: \text{It is morally permissible to kill an unjust aggressor, if killing this person is}
\text{the only way to save your own life or the lives of those you are responsible for}
\text{protecting, and if you can do so without doing something else that you are}
\text{morally required not to do.}
\]

Since PSD entails (but is not entailed by) PSD*, those who accept PSD should accept
PSD*, as well. Unlike PSD, however, PSD* is compatible with views like Otsuka’s
according to which it is morally impermissible to kill innocent aggressors.

Return now to KILL BILL 2. Let us assume that Bill is an unjust aggressor; he is
trying, without good reason, to torture and kill Beatrix’s daughter, and is blameworthy for
doing so. Let us also assume that in killing Bill, Beatrix did not do anything else that she
was obviously morally required not to do—killing him did not require that she also kill
an innocent bystander, for example. Given these additional assumptions, both PSD and
PSD* imply that it was morally permissible for Beatrix to kill Bill, since doing so is, by
hypothesis, the only way to prevent him from torturing and killing her child. I assume
that if it was permissible for \( S \) to \( A \), then it was not objectively wrong for \( S \) to \( A \). Hence, it
was not objectively wrong for Beatrix to kill Bill. Since she is blameworthy for killing
him nonetheless, it follows that the Objective View is false.
Blaming someone for what he did seems to essentially involve a desire that the person not have performed the action in question.\footnote{For the view that blaming involves a desire that the person being blamed not have acted as he did, see Sher 2006.} If this is right, then in order to blame Beatrix for killing Bill, we must have a desire that she not kill him. Some of us, however, are no doubt glad that Beatrix killed Bill, for in doing this she (unwittingly) prevented him from torturing and killing her daughter. Thus, it might be argued, those of us who are glad that Beatrix killed Bill will lack the requisite desire that Beatrix not kill him, in which case we cannot blame her for doing so.

There are several things to say in response to this worry. The first thing to note is this: the fact that I cannot blame someone for what he did does not entail that he is not blameworthy for doing it. If the devil tempts me to do something I know is very bad and I succumb to the temptation, the devil may not be able to blame me for what I did, since he presumably lacks a desire that I not perform the action in question. It clearly does not follow, however, that I am not blameworthy for my action. So even if it is true that we are not able to blame Beatrix for killing Bill, because we lack a desire essential to blaming her for this action, it might nevertheless still be true that she is blameworthy for killing him.

A second thing to notice is that it is entirely possible to have opposing desires. I might desire to have ice cream right now (because it is incredibly delicious) and, at the same time, desire not to have ice cream right now (because I’m trying to watch my weight). Similarly, we might desire that Beatrix not kill Bill, if, for instance, we are concerned that she avoids subjective wrongdoing, and, at the same time, be very glad that she killed him. So even if we want Beatrix to kill Bill, so that the daughter will be saved, and are glad that she killed him, that does not necessarily preclude us from having an opposing desire that she not kill him.

One further point is worth emphasizing before moving on. Being glad \textit{that} somebody did something is evidently consistent with blaming him for doing it. Consider:
HOOKY: Earl’s parents tell him not to skip school today, but he disobeys them. Lucky for him and his family he did, though. There was an explosion at his school in which all of his teachers and classmates were killed.

If you were Earl’s parents, how would you feel about your son and his behavior in this case? If you are like me, you would probably have mixed emotions. On the one hand, you will doubtlessly be very glad that he behaved as he did, for if he hadn’t, he would almost certainly be dead. On the other hand, you might reasonably be upset with him for skipping school, since he thereby deliberately disobeyed you. For a similar reason, it might be appropriate for Earl to feel guilty for skipping school, even if he is ultimately glad that he did.

I suggest that we are in a similar situation when it comes to Beatrix’s behavior in Kill Bill 2. We might reasonably be upset with her for killing Bill, and she might reasonably feel guilty for doing so, insofar as she thereby freely did what she believed to be objectively wrong despite believing that she could avoid wrongdoing by not killing him. That does not change the fact that we might be glad that Beatrix killed Bill. Just as it would be appropriate for Earl’s parents to be displeased with him for playing hooky despite being glad that he did so, so too Beatrix can be blameworthy for killing Bill even though we might be glad that she killed him.

At this point, a proponent of the Objective View might concede that Beatrix is blameworthy, but might insist that she is not blameworthy for her actions per se but rather for her willingness to do something she regarded as objectively wrong.\(^{15}\) What are we to make of this suggestion? I’m not entirely sure how to answer that question, mainly because I’m not sure what willingness to engage in wrongdoing amounts to.

Suppose $S$ is deliberating about which of several mutually exclusive courses of action, $A_1 \ldots A_n$, to perform. One thing we might say is that $S$ is willing to engage in objective wrongdoing just in case there is at least one action, $A_i$, among $A_1 \ldots A_n$ such that (i) $S$ regards $A_i$ as objectively wrong, and (ii) $S$ remains motivationally open to performing $A_i$ nevertheless—i.e., $A_i$ is still a live option for $S$; it is something $S$ is considering doing despite believing that it is morally wrong. Suppose something like this

\(^{15}\) E.g., see Copp 2003.
is correct, and consider two people, both of whom are willing to do moral wrong in the sense just adumbrated, but who differ in what they actually decide to do. The first person overcomes the temptation to do what he regards as morally wrong, whereas the second does not. The second, let us suppose, freely succumbs to the temptation, despite believing that she could avoid wrongdoing. Both individuals are therefore willing to do wrong, but only the second translates her willingness into action. I should think that the second person is therefore blameworthy for something the first is not, for something she did. Beatrix is like this second person; not only is she willing to do something she thinks she shouldn’t, she also succumbs to the temptation to do it. Thus, I would say that although she may well be blameworthy for her willingness to engage in objective wrongdoing, she is, in addition, blameworthy for something the first person is not, for something she did. What? An obvious answer is killing Bill.

By killing Bill, Beatrix did something she thought she shouldn’t. Perhaps it will be argued that what Beatrix is really blameworthy for is violating her own moral principles or sinning against her conscience or something like that. I agree, of course, that, by killing Bill, Beatrix violated her own moral principles and thereby sinned against her conscience. I also agree that, insofar as she did this freely despite believing that she could avoid moral wrongdoing, she is blameworthy for doing so. However, this is of help to the proponent of the Objective View only if it was objectively wrong for Beatrix to violate her moral principles in this case. But why suppose it was?

I imagine someone might insist that it is always objectively wrong to violate your moral principles, that it is always wrong to do what you think is wrong. However, I see no reason whatsoever to accept this claim and, moreover, it is counterintuitive in the extreme. It would imply, for instance, that, in HIDING PLACE, because Corrie believes that it is always objectively wrong to lie, it was objectively wrong for her to lie to the Nazis about whether she was hiding Jewish refugees in her home. But that is implausible.

The same point can be illustrated by a similar, but even more striking, example.

MADMAN: A madman threatens to mercilessly torture and kill 10,000 innocent people if Kathy does not tell a white lie. Kathy knows that the madman will do
exactly what he says he will do if she does not comply, but she believes that it is always wrong to lie, no matter what.

If it is always wrong to do what you think is wrong, then it is objectively wrong for Kathy to tell a white lie in this case. This implication is unacceptable, however. Surely it is permissible to tell a white lie, even if by telling the lie you would thereby sin against your own conscience, if doing so is necessary to save 10,000 innocent people being tortured and killed.

Violating one’s moral principles is, I think, much like lying in the following respect. Generally speaking, lying is a wrong kind of act; it is something we normally shouldn’t do. Nevertheless, it is plausible that there can be circumstances in which it is permissible to lie, whatever Kant might have said. Similarly, violating one’s moral principles is, generally speaking, a wrong kind of act; it is something that we normally shouldn’t do. But as cases like HIDING PLACE and MADMAN illustrate, there might nevertheless be circumstances in which it is permissible to violate your own moral principles. I suggest that KILL BILL 2 is also one of those situations. So I agree that Beatrix is blameworthy for violating her moral principles by killing Bill, but, contrary to what the Objective View implies, given the extraordinary nature of the situation, it arguably was not objectively wrong for to do so. Hence, the Objective View is false.

4.4 The Subjective View

Consider next the Subjective View:

**Subjective View of Blameworthiness**: Necessarily, if \( S \) is blameworthy for \( \phi \)-ing, then it was subjectively wrong for \( S \) to \( \phi \).

Given the difficulties facing the Objective View, several philosophers have found versions of the Subjective View appealing.\(^{16}\) For instance, Zimmerman says that, in addition to a freedom (or voluntariness or control) requirement, “Some other condition,

\(^{16}\) Haji, Swinburne 1989 and Zimmerman each accept versions of the Subjective View.
presumably mental, must also be satisfied,” in order for someone to be morally blameworthy for what he did (1997b, p. 411). What is this further, mental condition? According to Zimmerman, it is “a belief on the agent’s part that he or she is doing something [objectively] morally wrong” (ibid, p. 418).

Whatever initial appeal the Subjective View may posses, however, it ultimately isn’t very plausible. One obvious objection to the View, at least as it is presently formulated, arises from the possibility of culpable ignorance. If $S$ does not believe the truth that he has an all things considered obligation not to $A$, then it is not subjectively wrong for $S$ to $A$. However, if $S$ is culpable for his ignorance, many believe he can be blameworthy for $A$-ing nonetheless.

Consider, for instance, a case of medical negligence.

**ALLERGY:** Doc the doctor fails to read his patient’s chart carefully and thus doesn’t notice that she is severely allergic to penicillin. He treats her infection with penicillin and, as a result, she goes into anaphylactic shock.

Let us assume that Doc did not believe that he had an all things considered obligation not to give his patient penicillin and thus that it was not subjectively wrong for him to do so. Given this assumption, the Subjective View implies that Doc cannot be blameworthy for giving his patient penicillin. No doubt some will find this implausible. They will insist that if Doc is culpable for his failure to read the patient’s chart carefully and for the resulting ignorance of her allergy, then he may very well be blameworthy for giving her penicillin and for the harmful consequences of his action. If this is right, then contrary to what the Subjective View implies, Doc can indeed be blameworthy for giving the patient penicillin, even though it was not subjectively wrong for him to do so.

Proponents of the Subjective View are not without a response to this sort of objection, however. One thing they can say is that ignorance of the pertinent moral facts always excuses. Those who hold this view insist that, although a person who is ignorant of the fact that he is doing something morally wrong may indeed be blameworthy, he is not blameworthy for what he does in ignorance, but rather for some prior action or omission, the moral status of which he was aware, that causally contributed to his
ignorance. So, for instance, proponents of this view would say that, in ALLERGY, although Doc may be blameworthy for his failure to carefully read the patient’s chart or perhaps for some previous action or omission that led to this failure, he is not blameworthy for giving her penicillin.17

A slightly different response is to restrict the Subjective View to direct blameworthiness. A person is directly blameworthy for his action provided that he is blameworthy for it, and his blameworthiness does not depend on or derive from his blameworthiness for some prior act or omission of his. By contrast, if a person is indirectly blameworthy for what he did, his blameworthiness for his action will depend on or be derived from his blameworthiness for some prior act or omission. Those who restrict the Subjective View to direct blameworthiness would say that, in ALLERGY, Doc may indeed be blameworthy for giving his patient penicillin, but his blameworthiness for this action derives from and is dependent on his blameworthiness for something he did or failed to do earlier, such as failing to read her chart carefully, and is thus indirect. Had he not been blameworthy for his failure to read the chart (or for some previous action or omissions that led to this failure), he wouldn’t have been blameworthy for his subsequent ignorance or for the harmful actions to which his ignorance led. So, although Doc may indeed be blameworthy for doing something that was not subjectively wrong, his blameworthiness for this action is indirect, at best. Cases like ALLERGY therefore do not pose problems for versions of the Subjective View restricted to direct blameworthiness.18

So perhaps proponents of the Subjective View can avoid objections based on the possibility of moral ignorance. However, there is a related, but more trenchant problem with the Subjective View that cannot be so easily circumvented.19 Consider a

MORAL NIHILIST: Mack is a moral nihilist. He has been convinced by philosophical arguments that all claims of the form “S has an all things considered moral obligation not to A” or “it is objectively wrong to A” are all false. One day he freely tells a lie about a colleague despite being aware that this will seriously

17 For a defense of this view, see Holly Smith 1983.
19 Al Mele first pointed out this sort of objection to me.
damage her career. His reason for doing so was to avoid a minor personal inconvenience.

Being a moral nihilist, Mack doesn’t believe that he was subject to a moral obligation not to lie about his colleague. Hence, it was not subjectively wrong for him to do so. The Subjective View therefore implies that Mack cannot be blameworthy for what he did. However, this seems implausible.

Mack’s moral nihilism arguably doesn’t get him off the hook. Because he was aware of the harmful nature of his action, Mack’s behavior in this instance reflects ill on him and is grounds for a debit on his moral record. It would, moreover, be appropriate for his colleague to resent him for lying about her and thereby damaging her career. Likewise, it would be appropriate for us to be indignant toward Mack for what he did and for him to feel guilty about what happened. Evidently, then, Mack is blameworthy for lying about his colleague and thereby damaging her career despite the fact that it was not subjectively wrong for him to do so. Hence, the Subjective View is false.

Proponents of the Subjective View cannot circumvent this objection by claiming that there is some prior act or omission, the moral status of which Mack was not ignorant, for which he is really blameworthy. Even if there is no such act or omission, Mack’s behavior in this case would still reflect badly on him, the blaming attitudes would still be appropriate, and thus Mack would still be blameworthy for what he did. For similar reasons, proponents of the Subjective View cannot circumvent the present objection by restricting their View to direct blameworthiness. It seems that Mack would still be (directly) blameworthy for what he did, even if there is no prior action or omission of his for which he is blameworthy from which his blameworthiness for his behavior in the present case can be derived. Proponents of the Subjective View are therefore committed to the implausible claim that, given Mack’s moral skepticism, he is not blameworthy for lying about his colleague and thereby damaging her career, even though he did so freely and was fully aware of the harmful nature of his action.
4.5 The Disjunctive View

I’ve have argued that both the Objective and Subjective views are unacceptable. I conclude with a few brief remarks in defense of the Disjunctive View:

**Disjunctive View of Blameworthiness:** Necessarily, if $S$ is blameworthy for $\phi$-ing, then it was either objectively or subjectively wrong for $S$ to $\phi$ (or both).

The first thing to notice is that the Disjunctive View is consistent with all the cases of blameworthiness we have discussed. It correctly allows that Beatrix (both in KILL BILL and KILL BILL 2) is blameworthy for killing Bill, that Doc is blameworthy for giving his patient penicillin in ALLERGY, and that Mack is blameworthy for lying about his colleague in MORAL NIHLIST. The Disjunctive View is also consistent with the intuitive thought that a person is blameworthy for what he did only if there is something morally untoward about his behavior. Both objectively wrong actions and subjectively wrong actions are, or, in the case of the latter, can be, morally objectionable, though in different ways. Objectively wrong actions are morally objectionable simply in virtue of being violations of moral requirements. Subjectively wrong actions can also be morally objectionable if, for example, the agent thereby fails to act on what he takes to be, and what in other circumstances would be, sufficient reason not to perform the action in question, despite believing that he could avoid wrongdoing entirely by not performing that action. Given the difficulties facing both the Objective and Subjective views, then, some version of the Disjunctive View strikes me as the most plausible account of our intuitions about the connection between blameworthiness and moral wrongdoing.
CHAPTER 5

ASSESSING THE ALTERNATIVES

In the preceding three chapters I examined the main extant arguments for PAP and found each of them wanting. Where then does that leave the debate over the principle? My aim in this final chapter is to address this question. I argue that, in the absence of any further considerations either for or against PAP, debate over the principle and the traditional argument for the incompatibility of determinism and responsibility in which it figures is a classic instance of what John Martin Fischer (1994) calls a “dialectical stalemate.”

5.1 Call It a Draw

Philosophical argument often relies on appeals to common usage, examples and intuitive judgments about those examples to establish certain premises and conclusions. Sometimes, however, we reach a point in the discussion at which further appeals to these things ceases to be effective in moving the discussion forward. In his book *The Metaphysics of Free Will* John Martin Fischer refers to these argumentative impasses as “dialectical stalemates.”

At what point has a debate reached a dialectical stalemate? Fischer’s answer to this question can be found in the following passage:

Frequently in philosophy we are engaged in considering a certain argument (or family of arguments) for some claim C. The argument employs a principle P. Allegedly, P supports C. Now the proponent of the argument may be called upon to support the principle, and he may do so by invoking a set of examples (or other considerations). Based on these examples (or other considerations), he argues that the principle and thus also the philosophical claim are to be accepted.

But the opponent of the argument may respond as follows. The examples are not sufficient to establish the principle P. One could embrace all the examples
and not adduce P to explain them: rather, it is alleged that a weaker principle, P*, is all that is decisively established by the examples (or other considerations). Further, P*, in contrast to P, does not support C. Finally, it is very hard to see how one could decisively establish P. One reason it is so difficult is that it at least appears that one cannot invoke a particular example which would decisively establish P without begging the question in a straightforward fashion against either the opponent of P or the opponent of C. Further, it also seems that one cannot invoke a particular example which would decisively refute P without begging the question against the proponent of P or the proponent of C. These conditions mark out a distinctive—and particularly precarious—spot in dialectical space. (1994, p. 83)

These remarks suggest the following set of conditions that together are jointly sufficient for a dialectical stalemate:

1. An opponent of P or C can embrace all the examples (or other considerations) cited in favor of P and yet can (plausibly) argue that a weaker principle, P*, which does not support C, is all that is decisively established by the examples (or other considerations).
2. One cannot invoke a particular example which would decisively establish P without begging the question in a straightforward fashion against either the opponent of P or the opponent of C.
3. One cannot invoke a particular example which would decisively refute P without begging the question against the proponent of P or the proponent of C.

To these three conditions, I think we should add a fourth:

4. There are no further compelling arguments for or against P.
The reason for adding condition 4 is that there might be arguments in favor of P the premises of which cannot all be accepted by those who reject P (e.g., because together these premises entail P). Such arguments are not covered by condition 1.

I will assume that any philosophical dispute that satisfies these four conditions has reached a dialectical stalemate. Now, let P = PAP and let C = incompatibilism about determinism and responsibility. I suggest that the debate over PAP, and hence the debate over the traditional argument for the incompatibility of determinism and moral responsibility introduced in chapter one, has reached this unenviable position in dialectical space. As things now stand, these disputes end in a draw. To see this, let’s consider each condition in turn.

Begin with condition 1. The sort of examples often used by proponents of PAP to illustrate the plausibility of the principle can, it seems to me, be explained without appealing to that principle. These examples typically involve agents that act as a result of covert manipulation or irresistible impulses caused by drugs or various psychological disorders. Consider, for instance, an example introduced in chapter one.

**MORNING BREW:** A maniacal villain has covertly slipped a powerful, mind altering drug into Al’s morning coffee that produces in him an irresistible urge to strangle his friend, Bob. Powerless to resist the urge, Al rushes over to Bob’s house and strangles him.

Al is clearly not blameworthy for strangling Bob, and although one might explain this fact by appealing to PAP, it could also be explained by appealing to other considerations instead. For instance, it might plausibly be argued that, owing to the influence of the drug, Al is not appropriately responsive to the moral reasons against strangling Bob, and that a person must be suitably responsive to such reasons in order to be blameworthy for his actions.¹ I suspect that similar remarks could be made, *mutatis mutandis*, for other examples typically cited in support of PAP. Evidently, then, condition 1 is satisfied.

¹ This sort of explanation for cases like MORNING BREW is developed by Fischer and Ravizza 1998.

88
Conditions 2 seems to be satisfied, as well. It is hard to see how any particular example could establish the truth of a general principle like PAP without begging the question against opponents of the principle or opponents of incompatibilism.

What about condition 3? Here things are a bit more controversial, but it nevertheless seems to me that this third condition is also satisfied. The main challenges to PAP are those based on Frankfurt-style cases. However, in my estimation those challenges can be met. As I explained in chapters one and two, there are various places at which arguments against PAP based on Frankfurt-style cases can be challenged, and more importantly, I think, defenders of PAP have ways of explaining our intuitions about Frankfurt-style cases that are consistent with the likes of PAP. Moreover, it is hard to see how such examples could be further exploited to question the principle without begging the questions at issue.

Finally, what about condition 4? If what I have said in chapters two, three and four is correct, then the central lines of argument in favor of PAP fail to provide compelling reasons to accept the principle. Hence, in the absence of any further considerations either for or against PAP, I think we can assume that condition 4 is presently satisfied.

My use of the word ‘presently’ in the preceding sentence merits brief attention. It is perhaps tempting to treat dialectical stalemates as if they were static. On this way of thinking about these sorts of situations, once a dispute has reached this sort of impasse, all is said and done; nothing more can be said one way or the other, at least as far as philosophical argument and reflection on the issue is concerned. But this is not how we should think of the situation. As Fischer explains, dialectical stalemates needn’t result in philosophical despair. He insists that

Considerable philosophical ingenuity can be displayed in generating examples which invite one to accept the stronger principle [i.e., P] as well as the weaker principle [P*], or in explaining in a non-ad-hoc fashion exactly why one should only accept the weaker principle. Alternatively, philosophical creativity can issue in a restructuring of the problem; that is, one might find some other principle P*
which can be employed to establish C, or perhaps one can show in some way that C wasn’t that interesting after all. (1994, p. 84).

For instance, there might be ways of restructuring the traditional argument for incompatibilism that does not rely on PAP. Alternatively, a proponent of PAP might argue that, in order to have been suitably responsive to the moral considerations that bear on his actions, the agent must have been able to avoid doing what he did. Whether or not these strategies are promising, the point is that dialectical stalemates do not mean the end of the discussion. They should, however, causes us to rethink our approach to it.

My claim, then, is not that there is no way of resolving the debate over PAP and the traditional argument for incompatibilism (although that may in fact be the case), but rather that these debates are currently at an impasse. To break the impasse new considerations and approaches must be brought to bear on these issues. That, I suggest, is the task for those of us who continue to think about the relationship between determinism, moral responsibility and alternative possibilities.

5.2 The Traditional Argument for Incompatibilism

The debate between compatibilists and incompatibilists has been in the background for much of this dissertation. Let me briefly touch on it explicitly. In chapter one I introduced the following traditional argument for the incompatibility of determinism and moral responsibility.

1. Necessarily, if determinism is true, then no one can do otherwise.
2. Necessarily, if no one can do otherwise, then no one is morally responsible for his actions (from PAP).
3. Therefore, necessarily, if determinism is true, then no one is ever morally responsible for his actions (from 1 and 2).

---

2 For two attempts by incompatibilists to do this, see Lamb 1977 and Speak 2007.
This argument, or some variation thereof, has long been at the center of the debate between compatibilists and incompatibilists, and PAP plays an important role in it. What are the implications for this argument if, as I contend, the debate over PAP has reached a dialectical stalemate?

In the preceding section I suggested that, because debate over one of the argument’s central premises (viz., PAP) has reached a stalemate, debate over the argument itself has reached an impasse as well. That seems to me correct. Hence, I do not think that this argument provides us with compelling reason to embrace incompatibilism about determinism and moral responsibility. Indeed, several incompatibilists have recognized the weakness of this traditional argument and have sought to argue for incompatibilism using alternative means.\(^3\) However, it does not follow from any of this that the traditional argument is of no use whatsoever to incompatibilists.

In a genuine dialectical stalemate of the sort Fischer describes, it may be that, barring any further considerations on behalf of P or C, proponents of the argument for C in which P figures have failed to provide us with a compelling argument for C. It does not follow, however, that the argument for C is of no use to its proponents. All arguments rest on assumptions, and at some point argument has to come to an end. There are some assumptions that cannot be justified by appealing to further considerations. Let’s call these \textit{basic} assumptions. A proponent of the argument for C might claim that P is a basic assumption, and assuming that P has not been refuted or is not otherwise implausible, a proponent of C might be rationally justified in accepting P even if he has no further argument in favor of that assumption.\(^4\) Even in a dialectical stalemate, then, a proponent of C might claim that an argument for C based on P establishes the rational acceptability

\(^3\) Among the alternative arguments for incompatibilism that have been proposed, the so-called Direct Argument and the Manipulation Argument are the two most prominent. The Direct Argument was first introduced by van Inwagen 1980. The Manipulation argument has been defended by a number of philosophers in recent years and comes in a variety of different forms. For a helpful discussion of the various kinds of manipulation arguments, see Mele 2008.

\(^4\) When saying that a person is “rationally justified” in believing some proposition \(p\), I mean simply to say that the person’s belief that \(p\) is epistemically aboveboard. Someone who is rationally justified in believing that \(p\) has a reasonable basis for accepting \(p\) and thus isn’t irrational for accepting \(p\). On this way of thinking about rational justification, one person might be rationally justified in accepting \(p\), whereas another person might be rationally justified in accepting \textit{not}-\(p\). Someone who grew up in a scientifically naive culture and who had no access to relevant evidence might be justified in believing that the sun revolves around the earth, whereas someone in our culture who has access to the relevant science would not be rationally justified in believing that.
of C, insofar as some people are justified in accepting P and C follows from P (together, perhaps, with other rationally acceptable premises).

Many proponents of PAP would insist that the principle is a basic assumption. For some, PAP is intuitively plausible. Assuming, then, that the principle has not been refuted by Frankfurt-style cases or other considerations, those who find PAP intuitively appealing might be rationally justified in accepting the principle even if it cannot be justified by appealing to further assumptions. In that case, assuming the rest of the traditional argument for incompatibilism is in order, these people might be rationally justified in accepting incompatibilism on the basis of the traditional argument. So when it comes to the traditional argument for the incompatibility of determinism and responsibility, although that argument may not provide compelling reason for the compatibilist or reasonable agnostic to embrace incompatibilism, it may nevertheless be of use in establishing the rational acceptability of incompatibilism.

5.3 Philosophical Progress

In this dissertation I have argued that the main arguments for PAP are unsuccessful and that debate over the principle and the argument for incompatibilism in which it figures is likely at an impasse. However, in defending these claims I am not saying that philosophical progress has not been made on these issues. Indeed, I think much progress has been made, despite the fact that no consensus on PAP has yet been reached.

It is perhaps tempting to think of philosophical progress along the following lines. One has made progress with respect to a certain thesis, T, if and only if one has decisively refuted T or, alternatively, if one has provided a compelling argument in favor of T, one that is capable of persuading opponents of T or at least reasonable agnostics about T. On this way of thinking about philosophical progress, such progress occurs if and only if T has been decisively refuted or vindicated. But this is probably not the best way to think about philosophical progress. As I mentioned in chapter one, there may be much to learn from studying arguments for (or against) T even if the debate over T itself is never resolved. This, I want to suggest, is certainly the case when it comes to the debate over PAP.
There are (at least) three important lessons to be learned, if the arguments of chapters two, three and four are correct. In chapter two, we saw that an initially appealing account of blame and blameworthiness, according to which blame invariably involves an expectation that the person being blamed not have acted as he did, does not have as much going for it as one might have initially thought. In chapter three, I defended the view that the notion of an action is an actual sequence notion. That is, whether something counts as an action depends on how the event was brought about and not on whether the agent had access to alternative possibilities. In chapter four, I examined the claim (which is central to one of the arguments for PAP) that blameworthiness requires objective wrongdoing, and argued that this claim is false. In its place I suggested a disjunctive view, according to which a person is blameworthy for what he did only if the action in question was either objectively or subjectively wrong. Each of these three results is philosophically interesting in its own right, quite independently of the debate over PAP. So even if, as I think, the arguments in favor of PAP do not succeed, and even if debate over the principle ends in a stalemate, I think those who insist that continued discussion of these issues is a waste of time are mistaken. We have learned, and I think will continue to learn, much by reflection on the connection (or lack thereof) between responsibility and alternative possibilities.
REFERENCES


BIографиЧнаЯ ПрактиЧна ЧРЯгча

Justin Capes grew up in Atlanta Georgia. He received his Ph.D. in philosophy from Florida State University, an M.A. in philosophy from Biola University and a B.A. in history (with an emphasis in the history of ideas) from Kennesaw State University. His primary research interests are in ethics and metaphysics. He is particularly interested in a number of interrelated issues at the intersection of these two fields concerning human agency, including: moral obligation, reasons for action, free will, moral responsibility and allied notions such as blame, desert, forgiveness and punishment. During his time at Florida State, he published articles on these issues in journals such as Philosophia and Philosophical Studies, and was a recipient of a 2011 Graduate Research and Creativity Award.