A NON-REDUCTIVE NATURALIST APPROACH TO MORAL EXPLANATION

by

Lei Zhong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2010

Doctoral Committee:
Professor Peter A. Railton, Chair
Professor Allan F. Gibbard
Professor Robert G. Pachella
Associate Professor Sarah Buss

2010

To my wife

ACKNOWLEDGEMENTS

First I must thank the faculty members of Michigan who have overseen my dissertation project. Peter Railton is not only a great philosopher, but also an unusually good dissertation chair. He always provides me with his sympathy, patience, encouragement, elegant comments, and vivid illustrations. His broad and strong background in many philosophical subfields helps me to develop my project as well as my general philosophical approach in a fruitfully intra-disciplinary way. His philosophical insights also contribute to the change of my metaethical view, from ethical non-naturalism (sometimes intuitionism, and sometimes Kantian constructivism) to naturalistic moral realism.

Allan Gibbard is a very original and sharp philosopher. He is always able to find the weakness of my views and challenges them incisively and convincingly. Owing to Allan's model of philosophical inquiry and his scrutiny of my work, I have been growing up together with the philosophical positions I hold along my academic journey.

I owe an immense debt to Stephen Darwall, although he left Michigan and hence is not on my committee. As a first-rate philosopher with great personality, Steve has placed an important influence on me. I have also learned a great deal from his non-consequentialist ethical approach and his incomparable specialty in the history of ethics. I appreciate very much his encouragement and support for me from the beginning of my philosophical career.

Sarah Buss is both a wonderful ethicist and a super nice person. I am grateful for her enthusiasm to communicate with me on my naturalistic-oriented work she personally disagrees. Sarah is full of curiosity and sincerity in philosophical discussion and communication; her philosophical inquiry is never colored by academic pretentiousness or any sleight of hand.

Robert Pachella is a talented cognitive psychologist and also has good background in philosophy. I have enjoyed much our inter-disciplinary conversations between philosophy and psychology. I am always moved by his enthusiastic defense of his philosophy hero, Donald Davidson. I also thank him for generously letting me use one of his offices for several years.

I would like to thank other professors of Michigan who contribute to my philosophical growth in some way or other: Elizabeth Anderson, Victor Caston, James Joyce, Eric Lormand, Louis Loeb, Lawrence Sklar, and Kendall Walton, and two outside philosophy professors from whom I have learned a lot: Mark Schroeder (University of Southern California) and Ted Warfield (University of Notre Dame).

I also want to thank the Michigan philosophy students who give me both academic help and moral support: Marie Jayasekera, Ivan Mayerhofer, Stephen Campbell, Aaron Bronfman, David Dick, Lina Jansson, Sam Liao, Dustin Locke, Eduardo Garcia-Ramirez, Amanda Roth, and Dustin Tucker.

Last, but not the least, I feel deeply grateful to the philosophy professors of Peking University who introduced me to the fantastic field of analytic philosophy and made my further philosophical pursuit possible: Jiaying Chen, Xiangdong Xu, Yongping Sun, Lian Cheng, Dunhua Zhao, Chuang Ye, and Linhe Han.

TABLE OF CONTENTS

ABSTRACT

Many philosophers insist that moral facts or properties play no role in explaining (non-normative) natural phenomena. The problem of moral explanation has raised metaphysical, semantic and epistemic challenges to contemporary moral realism. If supposed moral properties play no explanatory role whatsoever, how can they have legitimate metaphysical status? If alleged moral properties are impotent in explaining our usage of moral language, how can our moral terms refer to these properties at all? If putative moral facts are irrelevant to explaining the occurrence of moral beliefs, how can our moral beliefs be justified or warranted, and how can we have moral knowledge?

In my dissertation, I attempt to vindicate the explanatory efficacy of moral properties, while at the same time respecting the autonomy and normativity of morality. In doing so, I will advocate a sort of non-reductive ethical naturalism, according to which moral properties are *natural* properties (in the sense that they are open to empirical investigation), and yet remain irreducible to *non-normative* natural properties, such as psychological, biological, and sociological properties.

More specifically, I develop a form of *moral functionalism* to vindicate moral explanation. Moral functionalism understands moral properties as second-order, functional properties, the natures of which are characterized in terms of

functional roles. My version of moral functionalism has two particular features. First, it is a form of *a posteriori* moral functionalism. The moral theory used to characterize the functional roles of moral properties can only be discovered by appeal to empirical investigation (for this reason, my *a posteriori* version of moral functionalism is a sort of ethical *naturalism*). Second, it is a *holistic* version of moral functionalism. The functional role of a moral property cannot be identified solely in terms of non-normative properties; rather, it essentially involves a network of connections to *both* non-normative and normative properties (for this reason, my holistic version of moral functionalism is a sort of ethical *anti-reductionism*).

It is important to note that the question of whether moral properties have explanatory power bears on the question of what explanation or explanatoriness is. In my dissertation, I attempt to show that my vindication of moral explanation presupposes a plausible account of explanation. Obviously I cannot discuss all theories of explanation; my dissertation will rather focus on two influential accounts of explanation: the causalist model and the unificationist model. While the causalist model holds that to explain an event or state of affairs is to provide some information about its causal history, the unificationist model claims that explanation is a matter of providing a unified account of a range of different phenomena. Then I argue that moral properties understood as functional properties of the kind I have described can do causal-explanatory work *and* play a distinctive unifying role.

Chapter 1

Ethical Naturalism Meets Moral Explanation

The No-Explanation Argument against Moral Knowledge

One of the well-known debates in contemporary metaethics is the debate over moral explanation. This is a disagreement between what I call '*moral explanationists*', people who advocate the explanatory efficacy of moral facts or properties (Sturgeon 1985, 1998; Railton 1986, 1998; Boyd 1988; Sayre-McCord 1988; Brink 1989), and '*moral anti-explanationists*', those who deny the explanatory efficacy of the moral (Harman 1977, 1986; Blackburn 1991; Audi 1997; Thomson 1998; Leiter 2001; Gibbard 2003a).

This debate attracts much attention of moral philosophers, because the problem of moral explanation raises serious challenges to contemporary moral realism. Some challenges are metaphysical: if supposed moral properties play no explanatory role whatsoever, how can they have legitimate metaphysical status? Some challenges are semantic: if alleged moral properties are impotent in explaining our usage of moral language, how can our moral terms refer to these properties? And some challenges are epistemic: if putative moral facts are irrelevant to explaining descriptive phenomena, how can any moral convictions be justified, and how can we have moral knowledge?

Let me just focus on the epistemological challenge here. The so-called no-explanation argument against moral knowledge goes as follows:

(1) Moral beliefs can be knowledge only if moral facts or properties play an *irreplaceable*[1] role in explaining non-normative or *descriptive*[2] phenomena (the explanatory condition on moral knowledge).

(2) Putative moral facts or properties cannot play an irreplaceable role in explaining any descriptive phenomena (the explanatory inefficacy of moral properties).

(3) Therefore, no moral beliefs can be knowledge.

As a naturalistic moral realist, I accept (1). So, I attempt to reject (2) in my dissertation. Although arguing for (1) in enough detail is beyond the scope of my dissertation, I can say something about it here. In my view, (1) is based on two further premises:

(1*a*) A belief that P is *a posteriori* knowledge only if the fact that P can play an irreplaceable role in explaining descriptive phenomena.[3]

(1*b*) Moral knowledge can only be *a posteriori* knowledge.

(1) Therefore, moral beliefs can be knowledge only if moral facts or properties play an irreplaceable role in explaining descriptive phenomena.

(1*b*) is pretty reasonable for me, as I regard the moral field as a natural, empirically accessible area. I anticipate that ethical intuitionists (such as Moore and Audi) would deny (1*b*).[4] According to the ethical intuitionist, moral knowledge, like mathematical knowledge, is *a priori*. In their view, we can still have moral knowledge even if moral facts play no role in explaining any

[1] To say that moral properties play an *irreplaceable* role is to say that the explanatory role of moral properties is not *completely* excluded by the explanatory role of descriptive properties.
[2] In my whole dissertation, I use the term 'descriptive' in the sense of 'non-normative'.
[3] Some philosophers (such as Harman) think that the explanatory condition even applies to *a priori* knowledge. Although I personally disagree with this, such a view doesn't contradict my acceptance of the explanatory condition on moral knowledge.
[4] It seems some quasi-realists such as Gibbard also assert that we have *a priori* moral knowledge, although they understand moral truths in a minimalistic sense (see Gibbard 2003a).

descriptive phenomena, just as we can have mathematical knowledge even if mathematical entities and facts play no role in explaining natural phenomena. I have good reason to believe that the ethical non-naturalist's epistemology is highly dubious. If moral knowledge were *a priori*, it would be very hard to explain how we come to know any moral facts. But I don't plan to discuss this in any detail here; rather I will just assume that moral knowledge is *a posteriori*.

Premise (1*a*) is pretty convincing to me, too. Think about other kinds of *a posteriori* knowledge. It's reasonable to assume that we are justified in believing in the existence of atoms, protons, and so on. Why? It is partly because assuming the existence of these microphysical entities plays an irreplaceable role in explaining certain natural phenomena. Or take our ordinary empirical knowledge as an example. There is a table in the room, and then I have sensory experiences about the table, and finally I believe that there is a table in the room. It is reasonable to suppose that my belief is knowledge. It is partly because to assume that there is really a table explains my sensory experiences better than competing hypothesis (the hypothesis that I'm in dream, for example).

A Non-reductive Ethical Naturalist Proposal

Regardless of whether the explanatory impotence of moral facts entails the falsity of moral realism in general, it would after all raise a serious challenge to naturalistic moral realism:

> (1) Ethical naturalism essentially asserts that we have moral knowledge that is *a posteriori*.
>
> (2) A belief that P is *a posteriori* knowledge only if the fact that P can play an irreplaceable role in explaining descriptive phenomena.

(3) Putative moral facts or properties cannot play an irreplaceable role in explaining any descriptive phenomena.

(4) Therefore, ethical naturalism is false.

A popular response to the problem of moral explanation within the framework of ethical naturalism is the reductive ethical naturalist solution.[5] Reductive ethical naturalists argue that moral properties can have explanatory power because they are *reducible* to (or, in other words, turn out to be *identical* with) descriptive properties, such as psychological properties, sociological properties and biological properties. Since descriptive properties have explanatory power, moral properties, if reducible, would *ipso facto* have explanatory power (Harman 1986; Leiter 2001). However, it is very doubtful that moral facts or properties are ultimately nothing but purely descriptive facts or properties—ethical reductionism just buys the *reality* of morality at the cost of its *autonomy*. How can the property of being morally right, for example, turn out to be any psychological or sociological or biological property? At least, up to now no one has provided a convincing project to show that moral facts or properties are actually descriptive ones. Many philosophers have been skeptical about the possibility of such reductions, and I'm inclined to share their suspicions. What's more, as I will discuss later, there is a vicious circularity in the reductive ethical naturalist solution to moral explanation.

On the contrary, I will argue that *non-reductive* ethical naturalism can vindicate the explanatory efficacy of moral facts. That is, moral facts and

[5] Reductive ethical naturalism is the view that moral properties are reducible to, or identical with, descriptive natural properties. By contrast, non-reductive ethical naturalism claims that moral properties are natural properties, but yet remain irreducible to descriptive properties. I will discuss the definitions in more detail later.

properties can have explanatory power even though they are *irreducible* to descriptive properties. In Chapter 5 of my dissertation, I will propose a version of *moral functionalism*, to give a specific account of non-reductive ethical naturalism. *First*, my version of moral functionalism is a kind of second-order moral functionalism rather than first-order moral functionalism. According to second-order moral functionalism, moral properties are identified as the second-order properties of having some properties that play certain functional roles, but not as the first-order occupants or realizers that actually play those roles. *Second*, my moral functionalism is a form of synthetic moral functionalism, according to which the functional roles of moral properties can only be specified by appeal to empirical investigation, as opposed to analytic moral functionalism, the view that moral properties can be identified as certain functional properties solely by conceptual analysis. *Third*, my version of moral functionalism is a holistic version of moral functionalism, according to which the functional roles of moral properties cannot be captured without reference to other moral or normative properties. For these reasons, my version of moral functionalism is thus a sort of non-reductive ethical naturalism. If the no-explanation challenge is met in a framework of non-reductive ethical naturalism, then we can preserve the reality of morality without sacrificing its autonomy.

It is worth noticing that the reductive ethical naturalist approach doesn't need to discuss theories of explanation in the first place. *If* the reductive ethical naturalist were to show that moral properties are identical with descriptive properties, they would show that the explanatory power of moral properties is preserved, no matter what explanation is. In contrast, an anti-reductionist

approach (a non-reductive ethical naturalist approach, or a non-reductive physicalist approach) normally needs to discuss theories of explanation first, and then show whether the higher-order properties in question can have explanatory power beyond their base realizers according to these accounts of explanation.

So I need to show that my vindication of moral explanation presupposes a plausible account of explanation. Obviously I cannot discuss all theories of explanation; my dissertation will rather focus on two influential accounts of explanation: the causalist model and the unificationist model. While the causalist model holds that to explain an event or state of affairs is to provide some information about its causal history, the unificationist model claims that explanation is a matter of providing a unified account of a range of different phenomena. Thus, moral properties can have explanatory power only if they can have causal power or unifying power, according to the two models of explanation. Hence I will discuss the two models and their relevance to the topic of moral explanation.

The difference between the causal model and the unificationist model reveals a more general contrast between explanatory externalism and internalism (see Kim 1994).[6] According to explanatory externalism, an explanatory relation must be essentially constituted by some external, objective relation in the world (Salmon 1984; Lewis 1986). By contrast, according to the internalist, explanation is primarily an activity internal to an epistemic corpus: whether or not something is an explanation depends on factors internal to a body of knowledge (Friedman

[6] The internalism/externalism distinction is sometimes characterized in terms of epistemic vs. ontic notions of explanation.

1974; Kitcher 1989). In my view, explanation or explanatory relevance needs to be constituted by *both* ontic and epistemic factors. But I won't discuss this in my dissertation; rather, I will be neutral between different theories of explanation in arguing for the explanatory efficacy of moral properties.

The Definition of Non-Reductive Ethical Naturalism

Since I appeal to non-reductive ethical naturalism to save moral explanation, I should first clarify the notion of 'non-reductive ethical naturalism'. Roughly speaking, non-reductive ethical naturalism is the thesis that moral properties are themselves *natural* properties, but they are irreducible to (or non-identical with) other natural properties, i.e. *descriptive* natural properties. But I wish here to raise two questions concerning the definition of non-reductive ethical naturalism. I will discuss them respectively.

First, in what sense are moral properties *natural* properties? It is hard to give a universally-accepted definition of 'natural'. Here I adopt a popular usage, according to which to say "*x* is natural" is just to say "*x* is open to empirical investigation" (See Railton 1989; Kornblith 1994; Copp 2003; Kim 2003).

Some philosophers, such as Sturgeon, propose a different definition according to which x is natural if and only if it has causal power. But this definition fails to *conceptually* distinguish naturalism from non-naturalism or supernaturalism. Some non-naturalists (e.g. Moore) hold that non-natural properties can still have causal power. Also, if there are supernatural beings (God, angels, demons, etc.), they would, of course, have causal power. If we were

to define naturalness in terms of causal power, then ethical naturalism might overlap with other kinds of metaethical theories.

Someone might say that although it is *conceptually* possible that Moorean non-natural properties or supernatural entities have causal power, *as a matter of fact* Moorean non-natural properties are causally impotent and supernatural entities don't really exist (and hence have no causal power). So, naturalism can be distinguished from non-naturalism and supernaturalism. However, whether Moorean non-natural properties have causal power and whether supernatural beings exist are open questions. I don't want to rule out these options in a *definition* of naturalness. If non-natural properties and supernatural properties have no causal power, non-naturalism and supernaturalism would be regarded as false theories of naturalism on this definition rather than other metaphysical doctrines distinct from naturalism—but this is odd.

The crucial difference between naturalism and other views such as Moore's non-naturalism is rather that whereas naturalists claim that moral properties can be known only by empirical investigation, Moore maintains that moral properties (at least *basic* moral properties) can only be known *a priori* (i.e. without appealing to empirical investigation). Likewise, supernatural entities are standardly considered as beyond the scope of empirical investigation. If in the future we could use empirical methodology to discover the nature of ghosts (say), which are traditionally regarded as supernatural beings, then ghosts should no longer be classified into the category of supernatural beings. Thus, if we understand naturalness in the sense of 'being subject to empirical investigation',

then we can plausibly distinguish naturalism from non-naturalism and supernaturalism.

Second, according to non-reductive ethical naturalism, moral properties are not identical with descriptive (or non-normative) properties. But what does 'descriptive properties' mean? Of course, descriptive properties shouldn't be understood literally as properties that are NOT normative—otherwise, it would be trivially true that moral properties are irreducible to, or non-identical with, descriptive properties.

Rather, by 'descriptive properties', I mean properties that are investigated by descriptive disciplines like physics, biology, sociology, and psychology. Now we get the definition of non-reductive ethical naturalism:

> (Definition) Non-reductive ethical naturalism is the view that *a*) moral properties are natural properties; *b*) moral properties are not identical with any properties which are investigated by descriptive disciplines.

Some philosophers (e.g. Sturgeon and Jackson) adopt a slightly different definition of 'non-reductive ethical naturalism', according to which moral properties would be reducible to descriptive properties if moral properties can be signified or represented by descriptive terms. While this definition and my definition converge on most occasions, the two definitions may go apart in some cases, such as the case of holistic moral functionalism. Suppose that we can appeal to the Ramsification method to define holistic moral properties (by Ramsifying the moral theory that characterizes the functional network). Then we can signify moral properties by the Ramsey expressions, which contain no moral terms. Then, according to Sturgeon's definition, moral properties are reducible to descriptive properties. However, I very much doubt that this definition of 'non-

reductive' is of much philosophical interest. After all, in order to know what a moral property is, we must rely on the relations between this moral property and other moral properties within a functional network characterized by the moral theory. In other words, the understanding of any moral property is always parasitic upon the understandings of other moral or normative properties. According to my definition, since the properties picked out by Ramsey expressions may not be investigated by descriptive disciplines, the mere fact that moral properties are signified by Ramsey expressions doesn't establish a case of reducibility. In Chapter 5, I will discuss more my definition of non-reductive ethical naturalism and argue why my definition is better than the other definition.

The Explanation Challenge to Reductive Ethical Naturalism

Ethical naturalists generally believe that reductive ethical naturalism can help justify the explanatory efficacy of moral facts, although they disagree whether *only* reductive ethical naturalism can save moral explanation.[7]

But quite contrary to the prevailing view, I will argue that we cannot appeal to reductive ethical naturalism for saving moral explanation, as there is a circularity in the reductive naturalist solution. The so-called solution is this. First, the reductive ethical naturalist attempts to vindicate moral-descriptive reducibility or identity. Second, she appeals to the identity thesis to save moral

[7] Some philosophers argue that non-reductive ethical naturalism can also save moral explanation. See Sturgeon 1985; Brink 1989.

explanation. However, in what follows, I will show that the reductive naturalist project contains a vicious circularity.

To begin with, I want to ask: How can we show that moral properties are reducible to purely descriptive properties? As most contemporary ethical naturalists insist, we cannot establish the identity between moral properties and purely descriptive properties solely by *a priori* reasoning or conceptual analysis; instead, we can only discover the moral-descriptive identities, if they can be discovered at all, by empirical investigation.[8] But it seems that the only way to discover *a posteriori* identities is to argue that assuming such identities is an inference to the best explanation. Let's consider a case of *a posteriori* identity in science, say, "temperature = mean molecular kinetic energy". We have knowledge about temperature, and we also have knowledge about mean molecular kinetic energy. Moreover, we know that temperature and mean molecular kinetic energy always *co-occur*. Finally we infer as the best explanation that temperature is just mean molecular kinetic energy. Before we establish this identity, we already have (non-conceptual) knowledge about temperature and mean molecular kinetic energy. By the same token, it seems that we can show that goodness, for example, is *a posteriori* identical with pleasantness only if we already have (non-conceptual) knowledge about goodness (like "honesty is good", "lying is not good", etc). But according to ethical naturalism, the only way to argue that we have (non-conceptual) knowledge about goodness is to argue that the property of goodness has explanatory power.

[8] Here I assume that analytic ethical naturalism is conclusively defeated by Moore's open question argument.

So, there would be a circularity in the reductive naturalist solution: in order to show that moral properties or moral facts are explanatorily efficacious, she appeals to the thesis that moral properties are (*a posteriori*) reducible to purely descriptive properties, in order to show that moral properties are (*a posteriori*) reducible to purely descriptive properties, she needs to show that we have (non-conceptual) knowledge about moral properties, but in order to show that we have substantive moral knowledge, she needs to show that moral properties have explanatory power.

The reductive ethical naturalist might reply that there could be some way to get out of this circularity by rejecting the assumption that vindicating the explanatory power of moral facts requires justifying first the moral-descriptive identities, and the latter in turn requires vindicating first the explanatory power of moral facts. It may be suggested that we can vindicate the explanatory power of moral facts (or, vindicating the explanatory power of hypotheses that assume moral facts) and justify the moral-descriptive identities *simultaneously*. Let me put it this way. First, the reductive ethical naturalist can treat the normative moral theory that is alleged to state moral facts and the identity theory that is alleged to state putative moral-descriptive identities as merely working hypotheses rather than knowledge or justified theories. Then, the reductive ethical naturalist suggests, we apply the two hypotheses to our explanatory practice. If it turns out that the two hypotheses can help explain the phenomena better than competing theories, then we would be justified in holding both hypotheses. For example, we have a normative claim that keeping promise is morally right, and the identity thesis that the property of being right is the

property of maximizing the overall happiness. Then we get the inference that keeping promises will maximize the overall happiness. Since the inferred hypothesis probably has explanatory power, on this suggestion, we can vindicate the explanatory power of the putative moral fact and justify the identity claim at the same time in this way. To generalize, we can succeed in vindicating the explanatory power of moral facts and justifying the moral-descriptive identities without circularity.

However, this suggestion doesn't work. Call the normative ethical theory M and the identity theory T. We replace every moral term with a descriptive term according to theory T, and then we get the descriptive theory D. The above suggestion is actually this:

> (1) If D has explanatory power, then the conjunction of M and T would have explanatory power.
>
> (2) If the conjunction of M and T has explanatory power, then M would have explanatory power.
>
> (3) D has explanatory power.
>
> (4) *Therefore*, M has explanatory power.

I doubt that Premise (1) is tenable. It is important to note that the reason why the conjunction of M and T has explanatory power is just that the conjunction would entail D, which has explanatory power. So, why do we need M and T in the first place? We can simply appeal to the *simpler* theory D alone! It seems that the conjunction of M and T is explanatorily useless, and it follows that M is also explanatorily impotent.[9] Therefore, the above way to get out of the circularity is

[9] Of course, if moral properties were *in fact* descriptive properties, the explanatory power of moral properties couldn't be excluded by that of descriptive properties—they would have the same explanatory power. But in the current context, the identity theory T hasn't been justified to be

blocked. Then we can safely conclude that reductive ethical naturalism fails to help vindicate the explanatory efficacy of moral properties or moral facts.

What's even worse, it is not only the case that reductive ethical naturalism fails to help justify the explanatory efficacy of moral facts, but also the case that this failure to save moral explanation entails the falsity of reductive ethical naturalism itself. As I mentioned earlier, ethical naturalism is true only if moral facts have explanatory power. Now reductive ethical naturalism cannot vindicate the explanatory power of moral facts, and it follows that only non-reductive ethical naturalism can save moral explanation. Thus, the situation is this: either we fail to vindicate the explanatory power of moral properties, or succeed in saving moral explanation by assuming non-reductive ethical naturalism. On the one hand, if it turns out that moral properties have no explanatory power, then reductive ethical naturalism would be false. On the other hand, if we can save moral explanation but do so only by assuming that moral properties are irreducible to descriptive properties, then reductive ethical naturalism would be also doomed.

It is worth mentioning that my argument against reductive ethical naturalism won't prove too much. That is, it doesn't threaten reductionism in other areas. It is very possible that a macrophysical entity or property, for instance, turns out to be a microphysical entity or property, just as water is reducible to H_2O, or temperature is reducible to mean molecular kinetic energy. Why does reductionism have so different a fate in the moral domain? The reason

---

true; whether theory T is justified depends upon whether the conjunction of M and T has explanatory power.

is probably this. We have epistemic access to macrophysical objects *independent* of the access to microphysical objects. For example, we can know that there is water in the bottle without having any knowledge about the microphysical properties of water (including the knowledge that water is H_2O). Water plays an indispensable role in explaining our beliefs about water; the explanatory power of water is not excluded by that of H_2O. Then, according to our knowledge about H_2O and our knowledge about the correlations between water and H_2O, it could be finally shown that water is H_2O. Since we don't need to appeal to the micro-macro identities to vindicate the explanatory power of macrophysical entities and the knowledge about macrophysical entities, there is no circularity involved in establishing micro-macro identities.

But the moral-descriptive relations are different. Moral properties not only metaphysically rely upon descriptive properties, but also epistemically depend on the latter (Audi 1997; Zangwill 2005). Here 'epistemic dependence' is a technical term. To say moral properties epistemically depend upon descriptive properties is just to say that the knowledge of moral properties derives from the knowledge of descriptive properties (Audi 1997). When we make a moral judgment about an action, our doing so is based on our knowledge of the descriptive properties of the action. For example, I see some kids burning a cat. Then I have a moral belief or moral judgment "the action is wrong!" But why? It must be because the action has such and such descriptive properties, say, the property of burning a cat just for fun. There seems to be an obvious truth: if I judge an action is wrong, I must be committed to there being something that makes the action wrong. That is to say, an action is *m* because it is *d*; and if I judge an action to be *m* it must be the

case that I judge the action to be *m* because it is *d* (here *m* is a moral predicate and *d* is a descriptive predicate). Thus, given the epistemic dependence of the moral upon the descriptive, there is always a potential competition in explaining the occurrence of moral beliefs between hypotheses assuming only descriptive facts and hypotheses assuming further moral facts.[10] According to reductive ethical naturalism, the moral theory M is alleged to be co-extensive with, or reducible to, the descriptive theory D. If so, it seems that what M explains can be explained by D as well. What are the theoretical gains of assuming M in moral explanations? How can we justify the explanatory power of M, if it is reducible to D?

I wish to emphasize here that the explanatory impotence of M doesn't merely come from the epistemic dependence of the moral upon the descriptive, nor merely from the coextension of M and D, but from the combination. Consider two theories X and Y, which are concerned with two kinds of properties, *x* and *y*, respectively. For one thing, if *x* doesn't epistemically depend upon *y*, then *x* may have explanatory power even though X is coextensive with Y. For another thing, if X is not co-extensive with Y but holds a one-to-many relationship with Y, then *x* *may* have explanatory power, even though *x* is epistemically dependent upon *y*. In Chapter 3, 4, I will discuss the second point in more detail, and attempt to argue that multiply realizable (and hence irreducible[11]) moral properties can have

[10] This is different from the 'water' example, in which the person believes that there is water in the bottle not because she believes that there is H_2O in the bottle. So there is no such a competition in the explanation of her 'water' belief.

[11] In my terminology, if *a* is reducible to *b*, then necessarily *a* is identical with *b*. So, multiple realizability is a sufficient (perhaps not necessary) condition for irreducibility.

distinctive explanatory power in virtue of their *unifying* role, even though they epistemically depend upon descriptive base properties.

Gibbard argues that the explanatory role of moral concepts is excluded by the explanatory role of concepts in descriptive disciplines (Gibbard 2003b). But in my view, Gibbard's challenge only applies to reductive ethical naturalism. According to Gibbard, the properties signified by moral concepts are also signified by concepts in descriptive disciplines, and therefore the explanatory role of moral concepts can be always replaced or excluded by that of those descriptive concepts. However, if non-reductive ethical naturalism is true, moral properties cannot be signified by concepts in descriptive disciplines. Then the explanatory role of moral concepts may go beyond the explanatory role of those descriptive concepts.[12]

To sum up, if for something to be a natural property it must have explanatory power, and if the only way to vindicate the explanatory power of moral properties is to adopt the thesis of multiple realization, then reductive ethical naturalism would be false in any event. In the following chapters, I will aim to argue that non-reductive ethical naturalism can save moral explanation. If I'm right, ethical naturalism wouldn't fail in either form.

[12] It is worth noticing that Gibbard seems to hold the view that the explanatory role of *properties* is completely reducible to the explanatory role of *concepts*. As I understand explanation as involving both ontic and epistemic elements, I disagree with Gibbard on this.

Chapter 2

Harman against Moral Explanation

Harman's Problem

The discussion of moral explanation originated from Gilbert Harman's oft-cited book, *The Nature of Morality*. Harman attempts to argue that it is not reasonable to have any moral beliefs because putative moral facts are explanatorily irrelevant to the occurrences of our moral beliefs.

Almost all contemporary moral skeptics regard natural science as the paradigm of human knowledge. If their arguments against moral knowledge could be equally used to argue against the existence of scientific knowledge, then the argument is unpalatable, at best, or is reduced to absurdity, at worst. Some older arguments (such as the verificationist approach) fail to distinguish ethics from science, and for this reason these arguments lose their attractiveness. Harman insists that his argument can reject moral knowledge without threatening scientific knowledge, as there is a crucial distinction between moral explanation and scientific explanation. Let me now introduce Harman's view. For example, when you round a corner and see some children pour gasoline on a cat and ignite it, you say "that is wrong". It seems that your moral observation can be explained without reference to any moral facts (e.g. the fact that setting cats on

fire is wrong). Instead, we can explain your observation by citing completely descriptive facts, such as facts about the descriptive properties of the 'burning cat' act and facts about your psychology.

But scientific explanation *appears* to be quite different. Harman gives an example of scientific explanation. A physicist sees a vapor trail in a cloud chamber, and then says "there is a proton". Of course, we need to assume some non-physical facts, such as the fact about the physicist's psychology, in order to explain the physicist's observation. But it seems that we also need to assume the particular scientific fact that there is a proton in the cloud chamber that produces the vapor trail. Thus, Harman concludes that there is a fundamental distinction between science and ethics: scientific facts assumed by scientific theories can play a role in explaining our scientific beliefs, whereas moral facts assumed by moral theories play no role in the explanations of moral beliefs. Harman writes:

> You need to make assumptions about certain physical facts to explain the occurrence of the observations that support a scientific theory, but you do not seem to need to make assumptions about any moral facts to explain the occurrence of the so-called moral observations I have been talking about. In the moral case, it would seem that you need only make assumptions about the psychology or moral sensibility of the person making the moral observation. In the scientific case, theory is tested against the world (Harman 1977, p. 6).

It is important to note the way in which Harman is here using the term 'observation'. *First*, Harman doesn't confine observation to perceptual observation, but rather uses the term 'observation' in a broad sense. According to Harman, A observes that P if and only if A *non-inferentially* believes, or 'sees', that P. *Second*, and more important, Harman use the term 'observation' in the sense of the *occurrence* of observation. Harman distinguishes two different

senses of 'observation'. Let me use the above example again. You see some children set a cat on fire and immediately think, "That's wrong." In one sense, your observation is that what the children are doing is wrong. In another sense, your observation is your thinking that thought (Harman 1977, p. 8). Borrowing Sellars's helpful terminology, we should distinguish between the *observed* (the content of an observation, P) and the *observing* (the act or event of observing that P) and, similarly, between the *believed* and the *believing* (Sellars 1968). In my paper, I shall, following Harman, use the term 'observation' (and 'belief') in the second sense.

In Harman's view, moral principles or supposed moral facts might help explain observations in the first sense but not in the second sense.[13] Suppose that you make an observation that burning a cat is wrong. A general moral principle (say, the principle "causing pain to sentient beings just for fun is wrong") might well explain why burning a cat is wrong. However, Harman contends (and I agree) that a valid theory need explain not just what we observe, but also our observing what we observe. So, the arguments goes, moral theories are not justified in that moral facts are explanatorily irrelevant to the occurrences of moral beliefs.

Although Harman focuses on the explanation of moral beliefs, my dissertation discusses moral explanation in general, including not only explanations of moral beliefs, but also explanations of social practices and explanations of moral character. As long as we can show that moral facts play a

[13] I want to add something. In Harman's mind, to assume moral facts or moral properties can be part of a better explanation *only if* moral properties are reducible to other natural properties. But Harman seems to believe that moral properties cannot be reducible to descriptive natural properties.

role in explaining certain descriptive phenomena whatever it is, we may probably be justified in holding the corresponding moral beliefs—it doesn't matter much what kind of moral explanation we are talking about. Moreover, there are close connections among the different kinds of moral explanations. Whether moral facts are explanatorily relevant to moral beliefs may depend upon whether moral facts are explanatorily relevant to other sorts of phenomena. If we can show that assumed moral properties are explanatorily relevant to, say, social practices, then this gives us a prima facie reason that moral properties can be also explanatorily relevant to our moral beliefs. For example, if justice plays a distinctive role in explaining the stability and flourishing of a country, then we may be justified in assuming the existence of justice. Thus we can probably show that purported facts about justice are also explanatorily relevant to our beliefs about justice.

Some explanationists such as Nicholas Sturgeon complain that Harman adopts a double standard for moral explanation vs. scientific explanation: while Harman has assumed that scientific theory is roughly true in arguing for the explanatory relevance of scientific facts, he has rather assumed that ethical theory is general false in arguing against the explanatory relevance of moral facts. According to Sturgeon's interpretation, the reason why Harman regards putative moral facts as explanatorily impotent is precisely that Harman considers ethical theory as 'hopelessly mistaken' (Sturgeon 1985, p. 70).

Certainly, observation is always theory-laden. A physicist can't 'see' that a proton is passing through a cloud chamber without already knowing a lot of physics. Just because we assume that the background physical theory is 'at least roughly true', we believe that we must assume the fact about the proton in

explaining the physicist's scientific observation. Given the physical theory, if there is a vapor trail of the right kind in the cloud chamber, there must be a proton which causes the vapor trace. So, it follows that the physical fact about the proton does explain the observation. Sturgeon writes:

> I believe we should grant that, *if* the physicist's theory were somehow this badly mistaken, but all the macrophysical facts (including all the observable facts) were held fixed, then the physicist, since he does accept that theory, would still draw all the same conclusions that he actually does… But no skeptical conclusions follows; nor can Harman, since he does not intend to be a skeptic about physics, think they do… This looks like a reasonable explanation, of course, only on the assumption that the physicist's theory is at least roughly true (Sturgeon 1985, pp. 70-1).

But likewise if the ethical theory is roughly correct, then someone who is burning a cat just for fun is probably doing something wrong. Then we can also say that the moral fact in question plays a role in explaining the moral observation. Moreover, Harman wants to use the no-explanation argument to establish the skeptical thesis that moral theories are not knowledge. So *if* Harman appeals to the to-be-proven skeptical conclusion in order to argue that moral facts are explanatorily irrelevant, then he would beg the question—or, at least, Harman's no-explanation argument wouldn't be an *independent* argument for moral skepticism.

However, I feel that Sturgeon seems to misinterpret Harman's view here. It is important to note that Harman doesn't assume the non-existence of moral facts in arguing against the explanatory impotence of moral facts (see Harman 1986). His point is rather that *even if* moral facts obtain, they are still explanatorily irrelevant to moral observations. So, there may be other grounds for Harman's anti-explanationist position.

Harman's Counterfactual Test

Someone may ask: "On what grounds does Harman claim that moral facts have no explanatory power?" On Sturgeon's interpretation, Harman seems to offer a 'counterfactual test' to determine whether moral facts are explanatorily relevant (Sturgeon 1985, p. 66). Although Harman doesn't *explicitly* put forward the counterfactual test, I agree with Sturgeon that it is implied in Harman's theory (see Harman 1977, 1986).

Let's go back to the 'burning a cat' example. Mary sees some kids putting a cat on fire, and she makes a moral observation that this is wrong. Then, Harman seems to ask this question: Would Mary have reacted in just the same way, with the thought that the action is wrong, even if what the kids were doing *hadn't* been wrong, and could we explain Mary's reaction equally well on this assumption? It is important to note that the formulation of the counterfactual question is not very clear and there are two ways to understand the question. How to understand the counterfactual question depends upon how to understand the phrase 'what the kids were doing'. Is it a rigid designator or non-rigid description? If we use the phrase in a 'referential' rather than 'attributive' way, then the phrase would behave as a rigid designator which picks out the action of burning a cat in every possible world. Then the counterfactual question would be actually this: Would Mary have reacted in just the same way, with the thought that the action is wrong, even if the action of burning a cat hadn't been wrong? By contrast, if we use the phrase in an 'attributive' rather than 'referential' way, then the phrase would behave as a non-rigid, context-dependent description which may pick out different actions in different possible worlds. Imagine, for example, that in a

different possible world, those kids are burning a toy cat in front of Mary, which is a different action from that action in the actual world. Accordingly, we can get a different counterfactual question from the first one—I will discuss the second reading later.

According to Surgeon's reading, Harman seems to understand the counterfactual test for explanatory relevance in the first way: the moral fact that burning a cat is wrong would be explanatorily relevant to the corresponding moral belief only if the moral belief would not have occurred if burning a cat had not been wrong. To generalize, the moral fact that P is explanatorily relevant to the moral belief that P only if the moral belief that P would not have occurred if the moral fact that P had not obtained. But we would have the same moral belief, Harman maintains, even if the moral fact had not obtained—that is to say, the supposed moral fact fails the counterfactual test. Therefore, the argument runs, moral facts are explanatorily irrelevant to moral beliefs.

Since neither Harman nor Sturgeon explicitly offers an analysis of counterfactuals, I want to say something about the analysis of counterfactual conditionals. If the conditional "if A had not obtained, then B would not have obtained either" is understood as "in every possible world where A doesn't obtain, B doesn't obtain either", then the counterfactual test for explanatory efficacy would be overly strong. Almost all explanations which are widely regarded as legitimate would be ruled out by the counterfactual test. For example, I shoot someone and cause him to die. It is perfectly reasonable to say that my shooting is explanatorily relevant to this person's death. But we can totally imagine a possible world where I perform exactly the same action of shooting, but the

person somehow is not killed by me (perhaps because he wears a bullet-proof jacket). Thus, we need to resort to a better theory of counterfactuals. Here I adopt the standard analysis of counterfactuals proposed by David Lewis, according to which "if *A* had not occurred, then *B* would not have occurred" means that "some world where *A* doesn't occur and *B* doesn't occur either is closer to the actual world than is any world where *A* does not occur but *B* occurs" (Lewis 1973).

A General Worry about the Counterfactual Test

I have a general worry about the counterfactual test for explanatory efficacy, for the counterfactual test seems too demanding for explanatory relevance. According to the counterfactual test, fact A is relevant to the explanation of fact B only if B would not have obtained if A had not obtained. In other words, the counterfactual dependence of B upon A is a necessary condition for A's being explanatorily relevant to B. I will give an example to show that in some tricky cases, fact A can be relevant to the explanation of fact B even if B would have obtained as well if A had not obtained.

Consider a beautiful city tree. Because the city government doesn't want children to be disappointed in case the tree is stolen, the government set up a hologram projection mechanism to the effect that a hologram image of the tree would appear on the instant the real tree is removed.[14] Suppose that I dropped by the city and saw the tree on the city square. Then I made an observation "that is a tree". Nobody denies that the fact that there is a tree is relevant to the explanation of my observation (assuming that visual perception is a *reliable* way

[14] I owe this example to Peter Railton.

to discern trees in the actual world). But we can also imagine a counterfactual world in which the real tree is absent but there is a hologram image of the tree in front of me. Certainly I would have the same visual experience as the one in the actual world. So in the counterfactual situation I would form the same observational belief that there is a tree (Note: this counterfactual world is probably the *closest* possible world). Thus, according to Harman's counterfactual test, the fact that there is a tree is irrelevant to the explanation of my observation in the actual case. But this is obviously odd.

The example of hologram tree is a case of potential preemption, where if the actual cause had not occurred, then another potential cause would have occurred to bring about the effect. Given the phenomena of potential preemption, counterfactual dependence is not a necessary condition for explanatory relevance—that is to say, the counterfactual test fails.[15]

Sturgeon's Counterfactual Test

Even if the counterfactual test of explanatory relevance as such is valid, it seems that this test cannot be used to argue against the explanatory efficacy of *moral facts*. As we know, moral properties or facts supervene upon descriptive properties or facts.[16] In other words, two actions (characters, policies, etc.) cannot differ with respect to their moral properties without also differing with

[15] David Lewis has a similar concern when he appeals to the counterfactual dependence to characterize causation. This is why he regards counterfactual dependence as a sufficient rather than necessary condition for causation.

[16] In my paper, I adopt the strong version of supervenience, which is essentially this: A *supervenes* on B if and only if necessarily, if anything *x* has some property F in A, then there is at least one property G in B such that *x* has G, and necessarily everything that has G has F. See Kim 1984.

respect to their descriptive properties. Moreover, moral-descriptive supervenience is supposed to be metaphysically *necessary*—that is to say, it holds in every possible world.[17] According to the thesis of moral-descriptive supervenience, if two actions have the same descriptive natural properties, then necessarily, they will have the same moral property. So, if burning a cat is actually wrong, how can it be possible that an action having exactly the same natural properties is no longer wrong in a counterfactual world? Since there is no possible world where burning a cat is not wrong, it doesn't make sense to say that some impossible worlds are closer to the actual world than other impossible worlds. Thus, it seems that Harman's counterfactual test cannot work.

Someone might propose that we revise Harman's counterfactual test into a non-counterfactual version so as to retain Harman's original gist without inviting this criticism. The former version is: fact A is irrelevant to the explanation of fact B if B would have been explained just as well *even if A had not obtained*. Now we can change it to a non-counterfactual test: fact A is irrelevant to the explanation of fact B if B would be explained just as well *even if we don't invoke any theory assuming A* (Sayre-McCord 1988). In other words, moral facts have no explanatory power because the hypothesis assuming the existence of moral facts fails to provide a better explanation for moral observations than other hypotheses that don't assume this. However, this non-counterfactual test doesn't provide a further reason why moral facts are explanatorily impotent—whether moral facts

[17] Some noncognitivists, such as Blackburn and Gibbard, regard the thesis of moral-natural supervenience as merely a *conceptual* truth rather than a *metaphysical* claim. But I gather they don't deny that *if* there are moral facts, then moral facts are necessary facts. Since in the current discussion we have assumed that moral facts actually obtain, it is reasonable to regard moral-natural supervenience as a metaphysical relationship.

can figure in the best explanatory theories and play an irreplaceable explanatory role is exactly what is the debate at issue. So, if we use this modified test to argue against moral explanation, we have made no progress at all—although this test doesn't violate the thesis of moral-descriptive supervenience.

Now let's come back to Harman's counterfactual test. As I mentioned earlier, there are two different interpretations of the counterfactual question: "Would Mary have reacted in just the same way, with the thought that the action is wrong, even if what the kids were doing hadn't been wrong?" If we understand the phrase 'what the kids were doing' as a rigid designator which picks out the action of burning a cat in every possible world, then the counterfactual question would turn out to be Harman's question: "Would Mary have reacted in just the same way, with the thought that the action is wrong, even if burning a cat hadn't been wrong?" But if we regard the phrase as a non-rigid designator which can pick out different actions done by the kids in different possible worlds, then we will get a different counterfactual test, according to which the wrongness of the kids' action is explanatorily relevant to Mary's moral observation only if Mary would not have made the same observation if the kids had not done anything wrong. Then we need to determine whether some world where the kids are not doing anything wrong and hence Mary doesn't judge it wrong can be closer to the actual world than any possible world where the kids are not doing anything wrong but Mary still judges it wrong. The new counterfactual test is just what Sturgeon proposes. Sturgeon claims that his counterfactual test is more natural and reasonable than Harman's (Sturgeon 1986, pp. 66-7).

Due to moral-descriptive supervenience, in order to imagine the children not doing something wrong we have to suppose their action different from the actual one in some of its non-moral features as well. So Sturgeon asks us to imagine that the children are doing something different, not vastly different, just different enough for it not to be wrong. For example, they are burning an old toy cat. If you saw this, how would you react to it? So our counterfactual question becomes:

> Even if the children had been doing something else, something just different enough not to be wrong, would you have taken them even so to be doing something wrong?

In the situation, if you still think it is wrong (perhaps because you hate children enjoying themselves), then the moral property of wrongness would be irrelevant to the explanation of your reaction. But in most cases, people would say it is not wrong (let's assume this). So, some world where the kids are not doing anything wrong and hence Mary doesn't judge it wrong may be closer to the actual world than any possible world where the kids are not doing anything wrong but Mary still judges it wrong. Therefore, wrongness in the actual case can make a difference to your reaction: were they not doing something wrong, you would have reacted differently. To generalize, moral facts can be relevant in explaining moral observations.

I'm not sure whether I correctly understand how Sturgeon considers his counterfactual test. Does he consider the test as a sufficient condition for explanatory relevance, or just a necessary condition? I agree that Sturgeon's counterfactual test can be a necessary condition in normal cases (except the cases of causal preemption), but I doubt that it is a sufficient condition for explanatory

relevance. In other words, even if the explanations of moral observations pass Sturgeon's test, it doesn't follow that moral facts are explanatorily relevant.

Judith Thomson objects to Sturgeon's counterfactual test as insufficient for determining the explanatory efficacy of moral facts. She constructs an example in which although some fact is explanatorily inert, it would be considered as explanatorily potent by Sturgeon's counterfactual test. Imagine that Donald suddenly shouted "Boo!" in the middle of someone's speech. Surely the rudeness of Donald's behavior supervenes on its being an act of shouting "Boo!" during someone's speech. So if Donald's behavior had not been rude then it wouldn't have been an instance of shouting "Boo!" during someone's speech, and no sound would have appeared on a tape recording of the speech. And it is therefore true that:

> If Donald's behavior hadn't been rude, then no sound would have appeared on the tape.

But just as surely, Thomson argues, the fact that Donald's behavior was rude fails to explain the "Boo!" sound on the tape. The appearance of the "Boo!" sound is rather explained entirely by the shout itself (Thomson 1998, p. 218).

It is important to note that Thomson doesn't claim that the rudeness of Donald's behavior is explanatorily excluded by the shout *just because* the rudeness of Donald's behavior supervenes upon this shout—otherwise Thomson would prove too much, for by following the logic all higher-order properties (e.g. psychological properties, biological properties, etc.) would be regarded as explanatorily impotent. If an argument against moral explanation generalizes to threaten these other sorts of higher-order explanations, it is at risk of becoming a

reductio ad absurdum of moral anti-explanationism (Sturgeon 1998; Loeb 2005). Rather, Thomson's point is that Sturgeon's counterfactual test for explanatory relevance is insufficient, since the test can be passed in cases like Donald's, in which the supervening property seems paradigmatically epiphenomenal.

Thomson's Principle of Explanatory Relevance

It is one thing to say that a particular argument for a thesis is untenable, and quite another to say that the thesis itself is false. If one wants to argue against the explanatory relevance of moral facts, one needs to find a way to distinguish epiphenomenal from non-epiphenomenal supervening facts, and then argue that moral facts fall into the former group—similarly, if one wants to argue for the explanatory efficacy of moral facts, one also needs a way to make this distinction and then attempt to show that moral facts are in the latter category. In order to make such a distinction, Thomson seems to bring forward a general principle for explanatory relevance:

> If a person believes that F because he takes the fact that X to be a reason for believing that F, then F explains the person's belief only if F explains X (Thomson 1996).

Thomson gives an example to illustrate this. Suppose that Jones believes that the apple is red. And suppose that she believes that (F) the apple is red because (X) the apple appears red to Jones. Since the apple's being red explains why the apple appears red to Jones, the apple's being red can be said to explain Jones's belief, according to Thomson's principle. Moral explanations of moral beliefs, however, *seem* to fail this test on this principle. Suppose, for example, Mary believes that

(F) the kids are doing something wrong, because she takes the fact that (X) the kids are burning a cat to be a reason for believing that (F) the kids are doing something wrong. Does the fact that the kids are acting wrongly explain Mary's belief that the kids are acting wrongly? Not, according to Thomson's principle. For the fact that the kids are acting wrongly does not explain the fact that the kids are burning a cat—it doesn't make sense to say that supervenient facts explain subvenient facts. Quite on the contrary, the kids' burning a cat explains the fact that their action is wrong.

As I mentioned in Chapter 1, moral facts not only metaphysically supervene upon descriptive natural facts, but also epistemically depend on the latter. When one acquires a moral belief about an action, the belief must be based on her knowledge of the descriptive properties of the action. There seems to be an obvious truth: if one believes that an action is wrong, it must be the case that she believes that something makes the action wrong. That is to say, the descriptive fact upon which the moral fact supervenes in this case is an epistemic evidence for the person's moral belief. But since the moral fact cannot explain the subvenient, descriptive fact, the moral fact is thus explanatorily irrelevant to the person' moral belief, according to Thomson's principle. The argument goes as follows:

(1) The fact that F explains a person's belief that F only if the fact that F explains the evidence-fact that X.

(2) The evidence-facts for a moral belief that M must include some descriptive fact that N upon which the moral fact that M supervenes.

(3) So, the moral fact that M explains a person's belief that M only if the fact that M explains the fact that N.

(4) Any facts cannot explain their subvenient facts.

(5) So, M cannot explain N.

(6) Therefore, the fact that M fails to explain the belief that M.

It seems that moral explanations are different in this respect from other sorts of explanations that also appeal to higher-order facts. Whereas descriptive higher-order facts (such as biological facts, sociological facts, and geological facts) also supervene upon lower-order facts, they don't epistemically depend on the latter. In other words, lower-order facts are not always introduced as evidence for our beliefs about descriptive higher-order facts. For example, I have a belief that there is water in the bottle. And suppose that I have no knowledge of chemistry or physics at all. Although the fact that there is water supervenes upon certain chemical or physical facts (say, the fact that there is H_2O), such lower-order facts are not evidence for my belief that there is water. So, the fact that there is water is not required to explain its subvenient facts in order to explain my belief. This is, one may argue, why Thomson's principle won't rule out many higher-order facts as being explanatorily impotent. It seems that Thomson finds a way to distinguish epiphenomenal from non-epiphenomenal supervening facts in arguing against the explanatory efficacy of moral facts.

However, I want to indicate that Thomson's principle is problematic. I agree that lower-order facts are not always introduced as evidence for our beliefs about descriptive higher-order facts. Also, I agree that in the circumstances that lower-order facts are not introduced as evidence, higher-order facts can be explanatorily relevant to the beliefs about such facts, even though they supervene upon lower-order facts. But it doesn't follow that when lower-order facts *are*

introduced as evidence, higher-order facts must be explanatorily irrelevant to the beliefs about such facts. Let me give an example. Suppose that a neuroscientist Linda attempts to investigate the mental states of a patient Stephen. Linda found that C-fiber is firing in Stephen's brain. Based on her neuroscientific knowledge, Linda infers that Stephen is in pain. It is perfectly reasonable to say that the fact that Stephen is in pain explains Linda's belief that Stephen is in pain. But since the fact that Stephen is in pain doesn't explain the evidence-fact that C-fiber is firing in Stephen's brain, the fact that Stephen is in pain wouldn't, according to Thomson's principle, explain Linda's corresponding belief. This is, however, absurd! It seems that Thomson fails to distinguish epiphenomenal from non-epiphenomenal supervening facts by appeal to her principle.

Someone might be quick to point out that Linda is justified in believing that the fact that Stephen is in pain is explanatorily relevant to her corresponding belief only because Linda is already justified in holding the general fact that everyone who undergoes C-fiber firing is in pain. That is, the explanatory power of the fact about Stephen's pain derives from the explanatory power of this general fact. But, where does this background neuroscientific knowledge come from? Well, we must first have independent knowledge about people's pain, then we observe that when people are in pain, there is always C-fiber firing in their brains, and finally we infer by induction that everyone will be in pain if she undergoes C-fiber firing. The etiology of the general neuroscientific knowledge seems to reveal that we have independent epistemic access to high-order facts about pain. Nevertheless, the moral case seems different, since moral facts epistemically depend upon descriptive facts. *If* the fact that the kids are acting

wrongly is explanatorily relevant to Mary's moral observation, the explanatory power of this particular fact must come from the general fact that everyone who is burning a cat for fun is doing something wrong. But how to justify Mary's *general* moral belief about burning a cat, given that moral facts epistemically depend upon descriptive facts?

My response is that if moral facts are multiply realized by descriptive facts, the hypotheses assuming moral facts could provide more unified explanations than alternative hypotheses, even if moral facts epistemically depend on descriptive facts. Consider injustice as an example. Guided by our most robust beliefs about justice, we regard many different configurations of social, political, and economical factors as 'unjust'. But such different configurations normally cause the *same* kind of phenomena, such as instability, revolution, general dissent. Thus, the appeal to injustice as a multiply realizable moral property will contribute to the success of explanatory practice by unifying its realizing properties. And if the explanatory power of multiply realizable moral facts is not excluded by that of descriptive facts, our background moral beliefs can be vindicated in this way. I will discuss this issue in detail in Chapter 4. But here we can say, at least, that Thomson's principle *alone* fails to argue against the explanatory efficacy of moral facts.

Anticipation

As I have argued, both Harman and Thomson fail to abolish moral explanation. If one wants to argue against the explanatory efficacy of moral facts, she must find other ways to do so. My suggestion is that we need to get into theories of

explanation to show in what sense moral facts have, or don't have, explanatory power. I very much doubt that the issue of moral explanation can be settled by appeal to any quick intuition or simple principle (as Harman, Sturgeon and Thomson do) without discussing theories of explanation first.

There are two influential accounts of explanation: the causalist model and the unificationist model. While the causalist model holds that to explain an event or state of affairs is to provide some information about its causal history, the unificationist model claims that explanation is a matter of providing a unified account of a range of different phenomena. In my view, both the moral explanationist and anti-explanationist need to discuss these theories of explanation for vindicating or rejecting moral explanation.

So, my discussion in the current chapter anticipates further discussion in next two chapters: Chapter 3 is concerned with the question of whether moral properties can have causal power and hence have explanatory power on the causal model of explanation, and Chapter 4 concerns the question of whether moral facts can effect a unifying role in moral explanations and hence have explanatory power on the unificationist model of explanation.

Chapter 3

Moral Explanation and Moral Causation

Introduction

As I discussed in the last chapter, we cannot resolve the debate over moral explanation without understanding what explanation or explanatory relevance is. In this chapter, I will discuss the causal account of explanation and attempt to argue that moral properties understood as multiply realizable natural properties can have causal-explanatory power.

Some philosophers hold that there is a close connection between explanation and causality. A (validly) explains B if A plays a role in causing the occurrence of B (at least in the circumstances that B is an individual event). According to David Lewis, all explanations of individual events are causal explanations: "to explain an event is to provide some information about its causal history" (see Lewis 1986; also see Salmon 1984). Hence someone can suggest that whether a fact or property is explanatorily epiphenomenal depends upon whether it is causally epiphenomenal. For example, the reason why the rudeness of Donald's behavior is explanatorily irrelevant to the appearance of the sound on the tape is precisely that the rudeness of Donald's behavior didn't *cause* the appearance of the sound on the tape. So, if it turns out that moral facts or

properties have no causal power, then it would be shown that they have no explanatory power either. The moral anti-explanationist can thus argue that moral properties are explanatorily irrelevant precisely because moral properties fail to cause descriptive phenomena or events. Harman writes:

> What's needed is some account of how the actual wrongness of Albert's action could help to explain Jane's disapproval of it. And we have to be able to believe in this account. We cannot just make something up, saying, for example, that the wrongness of the act affects the quality of the light reflected into Jane's eyes, causing her to react negatively. That would be an example of wrongness manifesting itself in moral claims, but it is not something we can believe in (Harman 1986, p. 63).

It may be worth noticing that the causalist account of explanation is overly restrictive and hence (mistakenly) rules out some legitimate explanations. For example, mathematical entities don't have causal power in bringing about natural phenomena, but they can still play some role in explaining those phenomena. Moreover, normally we can reasonably appeal to subvenient properties to explain the instantiation of their supervenient properties, but it is hard to say that base properties *cause* their supervenient properties to instantiate. However, for my purpose, I don't have to tackle these issues in my dissertation. First, I understand moral properties as natural properties. So it is irrelevant to my discussion of moral explanation whether non-natural properties like mathematical properties that don't have causal power can have explanatory power. Second, in the typical cases of putative moral explanation, moral facts and descriptive facts that are supposed to be explained by moral facts don't stand in a supervenience relationship to each other. For example, we appeal to injustice to explain revolution, or appeal to moral wrongness to explain the occurrence of wrongness-beliefs. The relationship between these explanans and explanandum are not the

one between base facts or properties and higher-order facts or properties. So it is irrelevant to our discussion whether relations that are not causal relations (e.g. the supervenience relation) can be explanatory relations.

Therefore, it seems that the moral explanationist still needs to vindicate the causal power of moral properties in order to vindicate the explanatory efficacy of the moral. In what follows, I will just assume provisionally that the causal theory of explanation is true, and then see whether it raises a real threat to the explanatory efficacy of moral properties. Although Harman is talking about whether moral facts are causally relevant to the occurrences of corresponding moral beliefs, I will discuss moral causation in general. For moral facts cause, if any, moral beliefs usually by causing other events or states of affairs. For example, Mr. Brooks, an evil person, often performed some crazy behaviors, say, killing people just for fun. A detective Tracy happened to observe those behaviors, and then formed a belief that Mr. Brooks is evil. Mr. Brooks's evil character is causally relevant to Tracy's moral belief only if his evil character is causally relevant to his behaviors in the first place. Thus, in what follows, I will discuss the causal power of moral properties in general, not only the causal power of moral properties for producing moral beliefs.

Before I discuss the problem of moral causation, I would like to state three assumptions in my discussion of causation. *First*, I assume that entities of various kinds, such as events, states of affairs, processes, facts, and properties, can be all causally efficacious or relevant. *Second*, I will assume that there is no sharp distinction between, in Dretske's term, a *triggering* cause and a *structural* cause (Dretske 1993). What we select as the cause of an event (or fact) is probably

pragmatic matter that depends upon our interests, purposes, and knowledge. *Third*, my discussion of moral causation is neutral to different theories of causation. I will discuss two predominant theories of causation: the counterfactual theory of causation and the probabilistic theory of causation, and attempt to show that moral properties can be causally efficacious on both of the two accounts of causation.

A Local Epiphenomenalist Approach

I will call people who deny the causal efficacy of moral properties 'moral epiphenomenalists'. If moral epiphenomenalists were just arguing against non-naturalists in the debate concerning moral causation, then their objection would make a lot of sense. For it is hard to show how non-natural properties can play any causal role in bringing about our beliefs. But moral epiphenomenalists seem to maintain further that if moral properties are irreducible to descriptive natural properties, then moral properties would have no causal power either (Harman 1986; Audi 1997). According to those philosophers, non-reductive ethical naturalism, in the same boat as ethical non-naturalism, fails to accommodate the causal power of moral properties. However, why does the question of whether moral properties are *sui generis* natural properties matter to the question of whether moral properties are causally impotent? Different moral epiphenomenalists adopt different approaches: some are global, and others are local. Global epiphenomenalists attempt to reject moral causation by appeal to some general arguments against the causal efficacy of irreducible higher-order properties (including not only moral properties, but also, for example, mental

properties). Local epiphenomenalists, such as Audi, say nothing about whether other higher-order properties, if irreducible, have causal power. Instead, local epiphenomenalists usually tend to introduce some *specific* cases about moral explanation and attempt to convince the readers that the reference to moral facts in causal explanations could be replaced without loss of explanatory power with reference to certain descriptive natural facts. I will discuss the two approaches respectively. In the current section, let's consider the local epiphenomenalist approach first.

Audi introduces the example of injustice to bolster his epiphenomenalist view. He writes:

> One cannot know (and normally would not even believe) that there is such injustice except through some kind of awareness of, say, government seizure of land, arbitrary curfews, and police brutality, where these are construed behaviorally in terms of, for example, soldiers' occupying farmland, clearing streets at night, and clubbing non-protesters. But these are just the sorts of non-moral factors, that in their own right, we suppose… can perfectly well explain a revolt. They also seem to have causal power in a quite intuitive sense (Audi 1997, p. 118).

Audi's argument actually has two steps. First, he specifies the descriptive base of moral properties in his example. Second, he asks us to compare the two competing explanations that appeal to moral properties and descriptive properties respectively, and urges us to share his intuition that the putative causal power of moral properties would be totally excluded or replaced by the causal power of descriptive base properties.

Someone may doubt that we can always isolate and specify the descriptive natural properties upon which a moral property supervenes. In Audi's example, injustice is realized by government seizure of land, arbitrary curfews, and police

brutality, etc. It's easy to see that not all forms of government seizure of land are unjust. In some urgent circumstances, it may be just for the government to seize private lands. Perhaps Audi actually means that inappropriate government seizure of land is unjust. But what does 'inappropriate' mean? The concept seems to be a normative one. How about 'arbitrary curfews' or 'police brutality'? Like 'inappropriate', 'arbitrary' and 'brutality' are probably also moral or normative concepts. In a word, moral concepts and descriptive concepts are often entangled with each other. Even though moral properties supervene upon certain descriptive natural properties, we can not always specify the supervenience bases of moral properties. If this is true, then Audi's examples would fail to convince the reader that the putative causal power of moral properties are entirely excluded or replaced by the causal power of descriptive base properties.

Probably Audi would reply that moral properties not only metaphysically supervene upon descriptive properties, but also epistemically depend on the latter. This is, Audi would argue, why we can specify the descriptive bases of moral properties. When we make a moral judgment about an action, our doing so is based on our knowledge of the descriptive properties of the action. So, given the epistemic dependence of the moral upon the descriptive, there is always a potential competition between assuming moral facts and assuming only descriptive facts in explaining the occurrence of moral beliefs. For example, suppose that one has a moral belief that the action is M. And the reason for her is that the action is N. Then, we must determine which hypothesis is a better explanation of her belief that the action is M, the hypothesis assuming the moral

fact that the action is M, or the hypothesis assuming only the fact that the action is N. Audi writes:

> Whenever we explanatorily invoke a moral property, it will be in part on the basis of, or at least in the light of, some belief or presupposition to the effect that one or more natural properties is playing an explanatory role. We are thus in a position to rely—often unselfconsciously, for sure—on those other properties to do the explanatory work, and it is arguable that they, and not any moral property, are in fact what does it… . Our understanding of how the ascription of moral properties can explain (at least so far as causal explanation goes) seems wholly derivative from our understanding of how the relevant base properties can do so (Audi 1997, 118-9).

Strictly speaking, it is odd to say that (irreducible) moral properties are causally impotent *because* moral properties epistemically depend upon the base properties. For whether moral properties have causal power is a metaphysical issue and hence doesn't depend upon how we come to have beliefs about moral properties. But for the sake of charitable reading, I think Audi actually means that *we are not justified in believing that* moral properties have causal power, given that moral properties epistemically depend upon the base properties. Audi claims that the epistemic moral-descriptive dependence as a particular feature of morality explains why the best inference is that irreducible moral properties are causally impotent (whereas other irreducible higher-order properties may be not).

However, Audi's argument proves too much. Think about mental properties. Many philosophers, including Audi himself, agree that mental properties have causal power (Audi 1997). But how do we know about others' minds? Sure, we cannot directly access others' mental states; we can only infer others' mental states from our knowledge of, say, people's behavior, countenance,

and utterances, etc. Does it follow that the putative causal power of mental properties would be excluded by these states and events upon which mental states epistemically depend? It doesn't seem so. So I conclude that epistemic dependence is neither peculiar to the moral domain nor sufficient to show that moral properties are causally impotent. Thus, Audi still owes us an explanation of why moral properties have no causal power.

Someone might be quick to point out a disanalogy between psychological cases and moral cases. Certainly, the knowledge of moral properties is based on the knowledge of base descriptive properties, just as the knowledge of others' mind is based on the knowledge of others' behavior. But moral properties supervene upon, or are realized by, those descriptive properties, whereas people's mental states don't supervene upon their behavior or countenance (according most theories of mind). This is, the suggestion goes, why (irreducible) moral properties have no causal power, whereas mental properties can have.

However, if someone makes such a suggestion, she actually adopts a *global* (rather than local) metaphysical principle: if A supervenes upon, but is not identical with, B, then the causal power of A would be excluded by that of B. I will discuss the second, general approach and make some responses to it in the following sections.

A Global Epiphenomenalist Approach

In arguing against moral causation, some philosophers might appeal to some general approach to argue that the causal power of any irreducible higher-order properties including moral properties is excluded by the causal power of their

base properties. A global epiphenomenalist approach has some advantages over its local counterpart in that it doesn't have to specify the descriptive supervenience base of any moral property. Insofar as there *are* descriptive natural properties which realize moral properties, say, wrongness, the supervenience argument can be conjured up to argue against the causal power of wrongness, whatever the wrong-making property is. Although conceptually there may be different global epiphenomenalist arguments against moral causation, in my dissertation I will focus on Jaegwon Kim's 'supervenience argument', which is quite popular.

Kim contends that if higher-order properties *merely supervene upon* (supervene upon but are not identical with) their base properties, the putative causal power of higher-order properties would be excluded by their base properties (Kim 1998, 2005). By following this logic, the causal power of irreducible moral properties would be excluded by their descriptive base properties—but if moral properties were reducible to, or identical with, base properties, the causal power of moral properties may be preserved. Kim's supervenience argument presupposes what he calls the 'Principle of Causal Inheritance':

> If M is instantiated on a given occasion by being realized by P, then the causal powers of *this instance of* M are identical with (perhaps, a subset of) the causal powers of P. (Kim 1993, 208)

Putting the principle of causal inheritance together with other premises, we can have the supervenience argument:

> (P1) *Supervenience*: Higher-order properties (e.g. moral properties and mental properties) supervene on base properties. That is, if any system s instantiates a higher-order property M at *t*, there necessarily exists a base

property N such that s instantiates N at *t*, and necessarily anything instantiating M at any time instantiates N at that time.

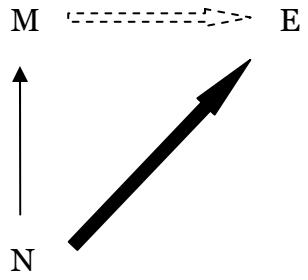
(P2) *Inheritance*: If M supervenes upon N, then the causal power of M would be identical with N.

(P3) *Non-overdetermination*: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.[18] And there is no systematic overdetermination in current cases.

(P4) *Irreducibility*: M is not identical with N.

(Conclusion) The causal power of M is excluded by that of N.

On the causal inheritance principle, since M supervenes upon N,[19] N should be regarded as a cause of E. Then given that M is irreducible to N and there is no causal overdetermination, it follows that the causal power of M is excluded by that of N. See the diagram below:



Let's consider the premises of the argument. The Supervenience Principle is widely-accepted, the denial of which would make moral-descriptive relationship extremely unintelligible. The Non-overdetermination Principle seems to be also pretty reasonable and is a standard view in contemporary metaphysics of

[18] Causal overdetermination is when an effect has more than one cause, and each cause would have brought about the effect if the other one had not done so. Consider the shootings of a tiger by two hunters. Suppose that the two shootings are set up in such a way that either would have killed the tiger if the other had failed. Then the death of the tiger is causally overdetermined by the two shootings. It is too *ad hoc* to suggest that generally M and N causally overdetermine the occurrence of E.

[19] Here N is assumed to be a minimally sufficient set of subvenient base properties of M. N is a minimally sufficient set of subvenient base properties of M if and only if there is no proper subset N' of N such that M supervenes upon N' when M supervenes upon N.

causation. Moreover, since I discuss whether irreducible moral properties can have causal power, I assume that the Irreducibility Principle is true for my purpose. Thus, the Causal Inheritance Principle is the only debatable thesis.

At first sight, the Causal Inheritance Principle is pretty intuitive. For example, someone dropped a glass cup on the wood floor, and then the cup was broken. Why was the cup broken? An apparent (partial) answer is: "Because the cup is fragile." But why is the cup fragile? It is because the cup has such and such molecular structure. Thus, the molecular structure that realizes fragility on this occasion appears to be a genuine cause of the cup's being broken. Granted that there is no overdetermination, if fragility is not identical with the molecular structure, then it seems to follow that the causal power of fragility would be excluded by the causal power of the molecular structure.

However, I want to indicate that the principle of causal inheritance is problematic. I will argue that even if A merely supervenes upon the base property (or property set) B, A can still have causal power (supposing that A is multiply realizable, rather than singly realizable, by the physical base B). And whether the supervenience argument is true depends on how we understand causation. In the following two sections, I will introduce two influential theories of causation, the counterfactual theory of causation and the probabilistic theory of causation, and attempt to show that the causal efficacy of irreducible moral properties won't be excluded by the causal power of their base properties according to the two theories of causation.

The Counterfactual Theory of Causation

In the current section, I will discuss the counterfactual theory of causation and attempt to show that the supervenience argument is problematic on a counterfactual account of causality.

According to the counterfactual theory of causation, where A and B are two distinct actual events, A causes B if B *counterfactually* depends upon A (i.e., if A had not occurred, then B would not have occurred either).[20] That is to say, *some* world where A doesn't occur and B doesn't occur either is closer to the actual world than is any world where A does not occur but B occurs (Lewis 1973).

In what follows, I will use some examples to illustrate my claim that even if higher-order properties merely supervene upon the base properties, they can still be causally efficacious. Let's consider an example of traffic light. Suppose that red lights have different shades of redness: some are scarlet, some are crimson.... Being scarlet is a particular way of being red; accordingly, the property of being scarlet is one of the base properties upon which the property of being red supervenes. And, assume that any kind of red light can cause the driver to stop: the driver sees a red traffic light, knows it means 'stop', and hence stops her car. Now consider a case in which a scarlet light causes the driver to stop. In this case, which property, the property of being scarlet or the property of being red, causes (or, more accurately, *quauses*[21]) the driver's stopping the car?

Let's discuss first whether the property of being scarlet causally contributes to the driver's stopping the car. Suppose that the property of being

[20] David Lewis regards counterfactual dependence as a sufficient condition rather than a necessary condition for causation.
[21] Property A quauses *e* if an instantiation of A causes *e* in virtue of (or *qua*) A. See Horgan 1989.

red, R, is multiply realized by base properties, such as the property of being scarlet, S, the property of being crimson, C, etc. It is reasonable to assume that any world where the light doesn't instantiate S and the driver doesn't stop the car is more remote from the actual world than a world where the light doesn't instantiate S but a slightly different shade of R, say, C (accordingly the light instantiates R) and the driver stops the car.[22] So, S is not a good candidate for the cause of the driver's stopping the car.

By contrast, a world where the light doesn't instantiate R and the driver doesn't stop the car could be closer to the actual world than any world where the light doesn't instantiate R but the driver still stops the car. Therefore, we should say that the property of being red is a better candidate for being the cause than the property of being scarlet in that the driver's stopping the car is counterfactually dependent upon the light's being red, but not the light's being scarlet.

Certainly, it doesn't follow from the *mere* fact that the driver's stopping the car doesn't counterfactually depend upon the light's being scarlet that the latter event is not the cause of the former event. But granted that there is no overdetermination in this case, since the driver's stopping the car counterfactually depends upon another event (i.e., the light's being red) which occurred at the same time and place as the light's being scarlet, it is reasonable to rule out the light's being scarlet as the cause of the driver's stopping the car.

[22] For example, a world where the light doesn't instantiate S and the driver doesn't stop the car is a world where the light instantiates the property of being green (say), so the driver doesn't stop the car, or a world where the light instantiates C (say) but mysteriously the driver doesn't stop the car.

In a word, insofar as the traffic light is red, the red light can play a role in causing the driver to stop, no matter which shade of red it is. The property of being red itself is enough to bring about the effect—the property of being scarlet seems irrelevant to the effect's being caused. In Stephen Yablo's term, the property of being scarlet is, in this case, 'casually superfluous' (Yablo 1992).

Another example. Consider a person, Ivan, who was in pain and hence walked to hospital to see doctor. Suppose that the property of having neurons firing in the anterior cortex (let's call the property 'P') is one realizer of Ivan's pain in this case. The relation between pain and property P is not a relation between the determinable and the determinate; rather, on some influential theories of mind, the relation between pain and P is a relation between a second-order, functional property and a first-order realizing property. Of course, pain can be realized in numerous different ways. Neuroscience suggests that other similar properties can also realize pain. For instance, when a person instantiates a property P' that is exactly like P except in which a particular firing neuron is different, she will also feel pain. Similarly, there could be countless such properties P'', P''', etc. In this example, which property, the property of being in pain, or the property of having neurons firing in the anterior cortex (property P), is causally relevant to Ivan's going to the hospital?

It is reasonable to assume that any world where Ivan doesn't instantiate P and Ivan doesn't go to the hospital is more remote from the actual world than a world where Ivan doesn't instantiate P but P' (accordingly Ivan feels pain) and Ivan goes to the hospital. So, P is not a good candidate for the cause of Ivan's going to the hospital. By contrast, a world where Ivan doesn't feel pain and he

doesn't go to the hospital could be closer to the actual world than any world where Ivan doesn't feel pain but he goes to the hospital. Therefore, we should say that the property of being in pain is a better candidate for being the cause than property P in that Ivan's going to the hospital is counterfactually dependent upon Ivan's feeling pain, but not Ivan's undergoing P. In this case, P is also casually superfluous just as is being scarlet in the example of traffic light.

The Probabilistic Theory of Causation

In this section I will discuss the probabilistic theory of causation, and argue that irreducible moral properties can have causal power on this account of causality. The central idea behind the theory is that causes raise the probabilities of their effects (Suppes 1970; Skyrms 1980; Elles 1991; Hitchcock 2002). Let me introduce the probabilistic principle of causation in a formal way:

> (P) C causes E (or, C is a positive causal factor for E) if and only if $P(E/C\&T_i) \geq P(E/\text{not-}C\&T_i)$ for all T_i, with strict inequality for at least one T_i, where T_i is a test situation consisting of a conjunction of independent causally relevant factors (Hitchcock 2002).

When such a conjunction of factors is conditioned on, those factors are said to be 'held fixed'. In order to determine whether C is a genuine cause of E, other *independent* causal factors need to be held fixed.[23] In this way we can do a better job to distinguish genuine causation from spurious correlations. For example,

[23] It is worth noticing that other causal factors which need to be held fixed must be independent causes of E. For example, let A and B be smoking and lung cancer, respectively. Suppose C is a causal intermediary, say the presence of tar in the lungs. If A causes B exclusively via C, then C will screen A off from B: given the presence (absence) of carcinogens in the lungs, the probability of lung cancer is not affected by whether those carcinogens got there by smoking. Thus we will not want to hold fixed any causes of B *that are themselves caused by A*. Let us call the set of all factors that are causes of B, but are not caused by A, the set of *independent* causes of B. See Hitchcock 2002.

suppose that smoking is a cause of heart disease, but exercise is an even stronger preventative of heart disease. And suppose that smoking is highly correlated with exercise: those who smoke are much more likely to exercise as well. Then it may be that smokers are, over all, less likely to suffer from heart disease than non-smokers. That is, letting A represent smoking, C exercise, and B heart disease, P(B/A) < P(B/not-A). However, if we conditionalize on whether one exercises or not, this inequality is reversed: P(B/A & C) > P(B/not-A & C), and P(B/A & not-C) > P(B/not-A & not-C) (see Hitchcock 2002).

For my purpose here, I don't need to discuss the subtle issues concerning the probabilistic account of causation. For the sake of brevity, I will use a simplified principle of probabilistic causation: C causes E (or, C is a positive causal factor for E) if and only if P(E/C) > P(E/not-C). My usage of this simplified principle won't affect our discussion of the supervenience argument. Accordingly, we can define neutral causal factor and negative causal factor as below:

C is a negative causal factor for E if and only if P(E/C) < P(E/not-C);

C is a neutral causal factor for E if and only if P(E/C) = P(E/not-C).

Now let's discuss the supervenience argument in the light of a probabilistic theory of causation. Consider the Thesis Q:

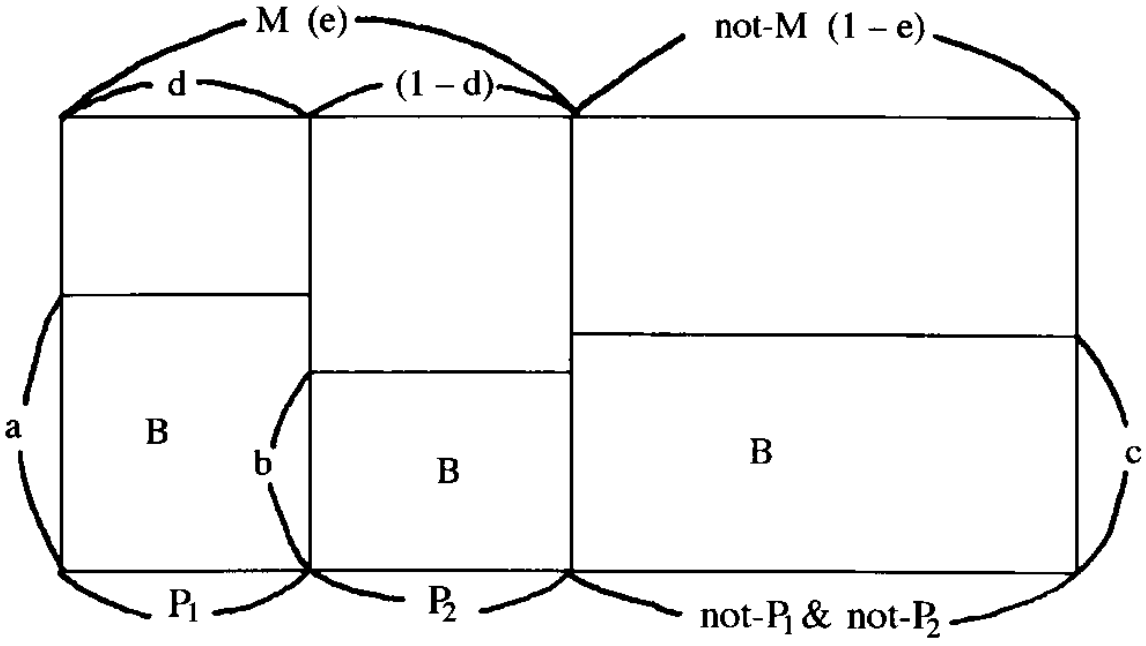
> (Q) If property M merely supervenes upon the base property (or property set) P, then the putative causal power of M will be excluded by that of P.

Suppose that M has only two realizers in the minimally sufficient base property set P, P_1 and P_2. Thesis Q says that in every circumstance that M is instantiated, the putative causal power of M would be excluded by that of P_1 or P_2. Thesis Q must entail:

(W) If M is a positive causal factor for effect E on an occasion, then P_i, a base property of M on this occasion, must be also a positive causal factor for E.

The reason is simple. If M is a positive causal factor for E and the base property P_i is not on the same occasion, then it is hard to say that the causal power of P_i would compete with, not to mention exclude, the causal power of M in this case.

In what follows, I will argue that thesis W is false. It is possible that M supervenes on P_i, and M is a positive causal factor for B, but P_i fails to be a positive causal factor for B. To see why, see the below figure that illustrate the relevant probabilities.[24]



Suppose that *M* has only two physical realizers, P_1 and P_2. Here $a = P(B/P_1)$, $b = P(B/P_2)$, $c = P(B/\text{not-}P_1\&\text{not-}P_2)$, $d = P(P_1/M)$, and $e = P(M)$. Let $a = 0.7$, $b \approx$

[24] This figure is from Kim 2000, p. 250.

0.416, c = 0.4, d = 0.5, and e = 0.1. Then we can get some interesting results as below:

$P(B/P_1) = 0.7 > P(B/\text{not-}P_1) \approx 0.4$;

$P(B/P_2) = P(B/\text{not-}P_2) \approx 0.416$;

$P(B/M) \approx 0.56 > P(B/\text{not-}M) = 0.4$.[25]

That is to say, whereas M is a positive causal factor for B, P_2 is a *neutral* causal factor for B, although P_1 is a positive causal factor for B (Kim 2000). Moreover, P2 can be a *negative* causal factor for B, while M is a positive causal factor for B. Let a = 0.8, b = 0.3, c = 0.4, d = 0.6, and e = 0.3. Then it can be shown that $P(B/M) > P(B/\text{not-}M)$, whereas $P(B/P_2) < P(B/\text{not-}P_2)$ (Kim 2000).

Furthermore, even if P_1 and P_2 are both positive causal factors for B, they can differ in their *quantitative* effects. That is to say, even if $P(B/P_1) > P(B/\text{not-}$

---

[25] First, let's get the values of some useful probabilities:
$P(M/P_1) = P(M/P_2) = 1$ (according to the supervenience relationship)
$P(P_1) = P(P_1\&M) = P(M)\ P(P_1/M) = e*d = 0.05$.
$P(\text{not-}P_1) = 1 - P(P_1) = 0.95$
$P(P_2) = P(P_2\&M) = P(M)\ P(P_2/M) = e*(1-d) = 0.05$.
$P(\text{not-}P_2) = 1 - P(P_2) = 0.95$
$P(\text{not-}P_1\&\text{not-}P_2) = P(\text{not-}M) = 1 - P(M) = 0.9$
Then we have proofs as below:
(1) $P(B/\text{not-}P_1)$
$= P(B/P_2)\ P(P_2/\text{not-}P_1) + P(B/\text{not-}P_1\&\text{not-}P_2)\ P(\text{not-}P_1\&\text{not-}P_2/\text{not-}P_1)$
$= P(B/P_2)\ [P(P_2\&\text{not-}P_1)/P(\text{not-}P_1)] + P(B/\text{not-}P_1\&\text{not-}P_2)\ [P((\text{not-}P_1\&\text{not-}P_2)\&\text{not-}P_1)/P(\text{not-}P_1)]$
$= P(B/P_2)\ [P(P_2)/P(\text{not-}P_1)] + P(B/\text{not-}P_1\&\text{not-}P_2)\ [P(\text{not-}P_1\&\text{not-}P_2)/P(\text{not-}P_1)]$
$= b*(0.05/0.95) + c*(0.9/0.95)$
$\approx 0.4 < P(B/P_1) = 0.7$
(2) $P(B/\text{not-}P_2)$
$= P(B/P_1)\ P(P_1/\text{not-}P_2) + P(B/\text{not-}P_1\&\text{not-}P_2)\ P(\text{not-}P_1\&\text{not-}P_2/\text{not-}P_2)$
$= P(B/P_1)\ [P(P_1\&\text{not-}P_2)/P(\text{not-}P_2)] + P(B/\text{not-}P_1\&\text{not-}P_2)\ [P((\text{not-}P_1\&\text{not-}P_2)\&\text{not-}P_2)/P(\text{not-}P_2)]$
$= P(B/P_1)\ [P(P_1)/P(\text{not-}P_2)] + P(B/\text{not-}P_1\&\text{not-}P_2)\ [P(\text{not-}P_1\&\text{not-}P_2)/P(\text{not-}P_2)]$
$= a*(0.05/0.95) + c*(0.9/0.95)$
$\approx 0.416 = P(B/P_2)$
(3) $P(B/M)$
$= P(B/P_1)\ P(P_1/M) + P(B/P_2)\ P(P_2/M)$
$= a*d + b*(1-d)$
$\approx 0.56 > P(B/\text{not-}M) = 0.4$

P_1), $P(B/P_2) > P(B/\text{not-}P_2)$, and $P(B/M) > P(B/\text{not-}M)$, it is possible that $P(B/M) > P(B/P_2)$, for example (See Kim 2000). Given that no causal overdetermination is involved in the circumstances, there is no option that P_2 and M are both causes of B. Then we must choose between P_2 and M for determining which is the cause of B. Since $P(B/M) > P(B/P_2)$, M is a better candidate for being the cause of B. Thus, it is not true that the causal power of irreducible higher-order properties would be excluded by their base properties on every occasion, on the probabilistic theory of causation.

Functionalist Causation and the Problem of Metaphysical Necessity

As I have argued in the proceeding two sections, the global epiphenomenalist approach is problematic, for the causal power of multiply realizable higher-order properties is not excluded by that of their base properties on both a counterfactual theory of causation and a probabilistic account of causation. Therefore, there is no reason why multiply realizable moral properties can't have causal power just like other higher-order properties.

Consider a compassionate person, Sam. Sam has a really kind heart to others, even to those who are not easy-going. He often makes an effort to help people for their own sake, even at a substantial cost to himself (see Cuneo 2006). Sam's compassion is realized by a complex, unified and stable constellation of psychological capacities, states and manifestations (let's call the realizing property base N). It is reasonable to assume that there could be countless similar properties (or property sets) N', N'', etc, which can also realize the virtue of compassion.

From our discussion of the counterfactual theory of causation, we can know that probably the virtue of compassion, rather than its realizer N, should be regarded as the cause of Sam's behavior, since Sam's behavior is counterfactually dependent upon Sam's virtue of compassion, but not the descriptive realizer N.

Also, moral virtues' being causally efficacious is totally compatible with a probabilistic account of causation. Let the virtue of compassion be C and physical behavior like Sam's B. From my earlier discussion of probabilistic causation, it is totally possible that P(B/C) > P(B/not-C), whereas P(B/N) = P(B/not-N), or even P(B/N) < P(B/not-N). Furthermore, even if P(B/N) > P(B/not-N), it is still possible that P(B/C) > P(B/N), and hence C should be better regarded as the cause of B than N. Whether the higher-order property is a normative property or a functional property seems irrelevant to whether it has causal power on the probabilistic account.

Since I advocate a version of moral functionalism, according to which moral properties are second-order, functional properties, I want to discuss functional causation in the rest of my chapter. Someone might insist that even if there are no general reasons against the causal power of multiply realizable higher-order properties, there are still particular reasons to show that the causal power of functional properties would be excluded by the causal power of its realizers—if so, that would be a problem for moral causation if moral properties are understood as functional properties.

In what follows, I will discuss a serious problem for functional causation, the so-called 'problem of the metaphysically necessary effects' (Shoemaker 1998; Rupert 2006). The putative problem is this. Functional properties are

individuated partly by their relation to the very effects those properties' instantiations are thought to cause. Consequently, functionalist causal generalizations would seem to have the following problematical structure: The state of being, among other things, a cause of *e* (under such-and-such conditions) causes *e* (under those conditions). Functional states in question are, by metaphysical necessity, causes of *e*; any state that does not cause *e* under the conditions is thereby a different kind of state. There is no such a possible world in which the functional state doesn't cause *e* under given conditions. Causal laws involving functional states and their effects are thus metaphysically necessary. But genuine causality should be contingent. (Rupert 2006, p. 256)

Let's take pain as an example again. Suppose that the functional characterization of pain includes its causing aversive behavior. Then attributing pain to a creature is partly to attribute to the creature some state or other that causes aversive behavior (under certain circumstances). Thus, to say that a pain state causes aversive behavior is just to say that a subject that is in some state or other that causes, among other things, aversive behavior causes aversive behavior. Such a causal 'law' appears vacuous, stating no more than that the cause of *e* causes *e*.

I put the problem of metaphysically necessary effects in a form way as below:

> (1) The alleged causal relation between a functional property P and its effect *e*, which is a causal output in the functional profile of P, is metaphysically necessary. That is, if P causes (or quauses) the occurrence of *e* under the conditions C, then there is no such a possible world in which P exists but fails to causes *e* under the same conditions C.

(2) But causation is metaphysically contingent. That is, if P actually causes *e* under the conditions C, then there is a possible world in which P exists but fails to causes *e* under the same conditions C.

(3) Therefore, P does not cause *e*.

Someone might indicate that premise (2) is controversial. According to a number of philosophers, causality is metaphysically necessary. For example, some philosophers hold a causal theory of properties, according to which properties are individuated in terms of their causal powers. Thus, on this view, causal relevance involving properties is metaphysically necessary. If property P is, by nature, causally relevant to the occurrence of *e* under conditions C, then there is no possible world in which a property is causally irrelevant to *e* under conditions C, and the property is P.

However, although premise (2) is not beyond question, I think it is pretty reasonable. For my purposes, I don't plan to discuss premise (2) in more detail. Instead, I will simply assume that premise (2) is true. In what follows, I will confine my discussion to premise (1), and see whether functional causation must involve metaphysical necessity.

It is worth noticing that whether functional causation is of metaphysical necessity is irrelevant to the difference between *a priori* (or analytic) functionalism and *a posteriori* (or synthetic) functionalism. That is to say, *if* this were a problem for functionalism *per se*, this would be also a problem for *a posteriori* functionalism. According to *a posteriori* functionalism, the functional roles of functional properties cannot be known by conceptual analysis or *a priori* reasoning alone; rather, the functional roles can only be discovered by empirical investigation. Nevertheless, *a posteriori* functionalism seems to be in the same

boat with analytic functionalism in facing the problem of metaphysically necessary effects. For example, suppose that by appeal to empirical investigation, we find that a property P is actually a functional property which is defined by functional role R. And suppose that functional role R involves the effect *e*. Then, to say that P causes *e* is just to say that the property of being in some state or other that causes *e* causes *e* (under certain conditions). If P had failed to cause *e* under right conditions, then P should not be regarded as the functional property characterized by functional role R. The causal generalization in question seems to be also metaphysically necessary, even though it is not conceptually necessary. The problem of metaphysically necessary effects still exists; whether the functional profile is discovered by empirical inquiry or conceptual analysis is beside the point (Rupert 2006).

Some functionalists, such as David Lewis, understand functional properties as first-order realizing properties. For Lewis and other like-minded philosophers, functional specifications are just used to fix the references of functional terms to the first-order realizers. For example, Lewis maintains that the term 'pain' actually picks out the first-order physical state that actually plays the 'pain' role (in relevant population and at a certain time). Suppose that C-fiber firing actually plays this role. Then pain is C-fiber firing, according to Lewis. If functional properties are first-order realizers as Lewis claims, functionalism of this kind wouldn't face the problem of metaphysically necessary effects. For it is not vacuous or metaphysically necessary that C-fiber firing causes aversive behavior (Rupert 2006). But as I will discuss in detail in Chapter 5, we have compelling reasons to adopt a second-order version of moral functionalism over

its first-order competitor. So this option is not available to me. I will argue that my second-order version of moral functionalism can still solve the problem of metaphysically necessary effects.

Before I introduce my responses, I want to discuss other philosophers' solution first. According to some philosophers, although a functional property P doesn't cause the effect *e*, which is involved in the causal-functional role R used to define P, P can still cause something else. Antony and Levine put it this way:

> [C]onsider the second-order property of dormativity. Dormativity is a property a substance possesses if and only if it possesses some first-order physical property that induces sleep… But arguably, the fact that this regularity ("Dormative substances cause sleep") holds does not certify the existence of distinct causal powers associated with dormativity…. However, consider this generalization: "dormative substances, if ingested before driving, cause traffic accidents." This generalization is true of anything that induces sleep, regardless of the mode of realization. It is an empirical, realization-independent regularity. (Antony and Levine 1997, pp. 11-2)

Since causing sleep is in the causal-functional profile which is used to characterize dormativity, it is not so much that dormativity *causes* sleep as that causing sleep (under right circumstances) is a *manifestation* of dormativity. But dormativity can cause something else which is not involved in its causal-functional profile, such as a traffic accident.

However, this solution doesn't really work. Note that the car accident is caused by the driver's falling asleep. And it cannot be suggested that the instantiation of dormativity and the instantiation of sleep overdetermine the car accident. So, dormativity can be causally relevant to the traffic accident only by being causally relevant to the driver's falling asleep. But Anthony and Levine

themselves concede that dormativity doesn't cause sleep. Then it follows that dormativity doesn't cause the car accident either.

Now I want to introduce my two responses. *First*, the characterizations of functional roles don't have to be all deterministic; rather, they can be probabilistic. For example, to say that a substance is dormative is, on a probabilistic view of functional roles, to say that the substance is in some state or other that is *likely* to cause someone to fall asleep under certain conditions. Then, it is not metaphysically necessary that the state that is likely to cause sleep (under certain conditions) causes sleep under the conditions. And let's assume that the probabilistic feature of functional roles is not just a matter of epistemic limitation, but a fundamental metaphysical fact. Even if we have full information, we still cannot give a deterministic characterization of dormativity (say). Moreover, even if functional properties denoted by ordinary terms like 'dormativity' and 'fragility' could be characterized deterministically, there might be some 'artificial' functional properties the characterizations of which involve probabilistic terms. For example, we can stipulate 'X' as 'the property of being in some state or other that is likely to cause sleep'. X is definitely a functional property, regardless of whether it is dormativity. If my above analysis is true, then it is far from conclusive that causality involving *any* functional property is of metaphysical necessity.

Second, I wish to stress that no matter whether the problem of metaphysically necessary effects is a challenge for the causal power of properties like pain and dormativity, the functional roles of which are purely causal roles, it is not a serious problem for functional causation in general. Not all functional

states are like pain and dormativity, and not all functional roles are purely causal roles. Consider the desire to be smart. According to functionalism, a mental state should be characterized in terms of the connections between this mental state and certain behaviors, dispositions and other mental states. A functionalist may offer a functionalization of this type of desire: *S* has a desire to be smart if and only if *S* would *tend* to do *x* (say, studying philosophy) when she *believes* that doing *x* is the best way to make her become smart…. But upon reflection, this suggestion is not right. For even if a person has a desire to be smart and she also believes that doing *x* is the best way to make her become smart, it is possible that she doesn't tend to do x. Perhaps we should say that *S* has a desire to be smart if and only if *S ought to* tend to do *x* when she believes that doing *x* is the best way to make her become smart. Here the connection between the desire to be smart and the behavior disposition to do x is not a causal connection.

The same is true of functionalist moral properties, the functional roles of which don't entirely consist of causal roles. As Jackson and Pettit put it,

> [F]unctionalism about the mind associates primarily causal roles with the various mental properties. By contrast, although the functionalist account of moral properties may include as part of the role definitive of fairness 'Many acts are right because they are fair', the 'because' is not a causal 'because'. The fairness of an action does not causally explain its rightness. (Jackson and Pettit 1995, p. 25)

Since the functional roles of moral properties don't entirely consist of causal connections, the problem of metaphysically necessary effects is not a serious problem for functionalist moral causation in particular.

Take rightness as an example. Rightness typically bears on the promotion of people's well-being. Rightness is intrinsically motivating. When one judges an

action to be right, she normally has a prima facie motivation to perform this action. An action's being right provides people with an important, even overridingly so, normative reason to do it.... We can appeal to this functional role to define what rightness is. The functional role doesn't consist of purely causal connections between rightness and other properties and states, etc. For example, the connection between a (potential) action's being right and the agent's tendency to do this action is not a causal connection; rather, it is the case that the agent *ought* to tend to do this action when she perceives the rightness of the action. On some respectable theories of causation, it may be shown that the rightness of an action causes a person to perform the action on an occasion. But there is no problem of metaphysically necessary effects. For it is not metaphysically necessary that the property of an action that makes it that the agent *ought to* do the action causes the agent to do it (whereas it may be metaphysically necessary that the property of an action that would cause the agent to do the action causes the agent to do it).

Conclusion

Given that the causal explanation requirement applies to the moral field, the moral explanationist needs to show that moral properties can be causal relevant to the occurrences of descriptive natural phenomena. In this chapter, I have discussed two influential theories of causation: the counterfactual theory of causation and the probabilistic theory of causation, and argued that irreducible moral properties could be causally efficacious on both of the two accounts of causation. Moreover, since I understand moral properties as functional

properties, I have also discussed functional causation and the putative problem of metaphysically necessary effects, which is alleged to be a problem for the causal efficacy of functional properties. Finally I have attempted to show that functional properties can have genuine causal power without metaphysically necessitating their effects.

Chapter 4

The Explanatory Unification of Moral Properties

Introduction

In the preceding chapter, I have argued that the causal power of irreducible moral properties is not excluded by that of their descriptive realizers. So, on a causal theory of explanation, moral facts or properties could have explanatory power.

However, we should be aware that the causal model of explanation is far from conclusive. There are many valid types of explanations that are not causal explanations at all, such as explanations appealing to mathematical entities, explanations involving the supervenience relationship (e.g. explanations of macrophysical qualities of an object in terms of its microphysical properties), and equilibrium explanations (explanations of the occurrence and persistence of an equilibrium state)[26].

The main competitor to the causal model is the unificationist account of explanation, which underscores the significance of unification for explanation or

[26] When we explain the occurrence and persistence of an equilibrium state, the actual initial conditions and the actual causal history about the equilibrium state seems explanatorily irrelevant. As Elliott Sober indicates, "where causal explanation shows how the event to be explained was in fact produced, equilibrium explanation shows how the event would have occurred regardless of which of a variety of causal scenarios actually transpired." See Sober 1983.

explanatory relevance. The difference between the causal model and the unificationist model reveals a more general contrast between explanatory externalism and internalism (see Kim 1994). The internalism/externalism distinction is sometimes characterized in terms of epistemic vs. ontic notions of explanation. According to explanatory externalism, an explanatory relation is essentially constituted by some external, objective relation in the world (Salmon 1984; Lewis 1986). By contrast, according to the explanatory internalist, explanation is primarily an activity internal to an epistemic corpus: whether or not something is an explanation depends on factors internal to a body of knowledge (Friedman 1974; Kitcher 1989).

For the purposes of my dissertation, in order to vindicate the explanatory efficacy of moral properties, my vindication should be compatible with different (major) theories of explanation. So, in this chapter, I will discuss a unificationist theory of explanation, and suggest how, on such a theory, multiply realizable moral properties can have explanatory power. In particular, I will propose a *local* and *soft* version of unificationism (in contrast to Kitcher's *global* and *hard* unificationism), and attempt to show that moral facts have explanatory power in virtue of their distinctive unifying role in explanations.

Kitcher's Unificationist Theory of Explanation

The basic idea of a unificationist account of explanation is that a good explanation normally provides a unified account of a range of different phenomena. Unificationism is attractive for at least two reasons. First, the unification model can easily accommodate certain types of explanations, such as

mathematical explanations and mereological explanations, which have proved problematic for competing theories (e.g. the causal-mechanical model). The reason why mathematical entities, for example, are explanatorily efficacious doesn't lie in their causal power, but in their unifying role. Second, the unification model promises to clarify the link between explanation and understanding. Successful unification, such as Newton's unification of terrestrial and celestial theories of motion, or Maxwell's unification of electricity and magnetism, exhibits connections or relationships between phenomena previously thought to be unrelated. A unifying explanation would advance our understanding of the world in this sense.

The unificationist account of explanation has been most successfully developed by Philip Kitcher. In this paper, I will propose a modest version of unificationism, which contrasts with Kitcher's strong unificationism. Let us begin our discussion with Kitcher's theory. According to Kitcher, the explanatory power of scientific theories, such as Newton's mechanics and Darwin's evolutionary biology, lies in the fact that they employ relatively few explanatory patterns to derive, and explain, large numbers of phenomena. Kitcher writes:

> Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same pattern of derivation again and again, and in demonstrating this, it teaches us how to reduce the number of facts we have to accept as ultimate (or brute) (Kitcher 1989, p. 432).

Kitcher's guiding idea is that explanation is a matter of deriving descriptions of many different phenomena by using as few 'argument patterns' (or explanatory patterns) as possible. In Kitcher's view, the fewer the patterns used, and the

greater the range of different conclusions derived, the more unified our explanations.

An *explanatory pattern* (or argument pattern) is an ordered triple consisting of a schematic argument, a set of filling instructions, and a classification of the schematic argument. A *schematic argument* is a sequence of schematic sentences. A *schematic sentence* is an expression obtained by replacing some, but not necessarily all, the nonlogical vocabulary occurring in a sentence with dummy letters. *Filling instructions* are directions for replacing the dummy letters of the schematic sentence such that, for each dummy letter, there is a direction that tells us how it should be replaced. A *classification* for a schematic argument is a set of statements describing which sentences in the schematic argument are premises and conclusions and what rules of inference are used (Kitcher 1989, p. 432).

Let's consider a simplified example of argument pattern for illustration.[27] Here is an argument, A:

(1)	Animals tend to experience pain at points of recent bodily damage.
(2)	Animals tend to wince when they experience pain.
(3)	Mike, an animal, was subject to recent bodily damage.
(C)	Mike winced.

First, we replace some, but not all, non-logical expressions by dummy letters, and get a schematic argument, SA:

[27] I borrow and modify this example from Franz-Peter Griesmaier's "Kitcher-style Unification and Explanatory Relevance", in *dialectica* 59 (2005).

(1')	S tends to experience pain at points of recent bodily damage.
(2')	S tends to do W when S experiences pain.
(3')	*p*, an S, was subject to recent bodily damage.
(C)	*p* does W.

Second, we specify the logical structure of this schematic argument by introducing a *classification*, C, which tells us which sentences are premises, which inference rules are applied, etc.:

Sentences (1'), (2'), and (3') are premises.
Sentence (C) is supported by the premises with a probability of r.
The rules are (*i*) probabilistic modus ponens and *(ii*) simplification.

Third, we specify a set of *filling instructions*, FI:

Replace 'S' by 'animals'.
Replace 'W' by 'wincing'.
Replace '*p*' by the name of an instance of S.

Further, Kitcher introduces the notion of '*explanatory store*' to develop his unificationist account. The *explanatory store*, E(K), is defined as the set of argument patterns that maximally unifies K, the set of statements endorsed by the scientific community. A particular explanatory pattern would count as a good or legitimate explanatory pattern if it belongs to the explanatory store E(K) (Kitcher 1989).

Here is a question: in order to best unify K, must the premises of argument patterns also be members of K, i.e. statements reasonably accepted as true? One may require that the premises of argument patterns be also members of K. On this view, to find the unifying theory for K, we first systematize K by choosing some elements of K as premises from which we can derive all other elements of K.

Since there will be generally more than one systematization, we need to choose the most unifying way of systematization.

However, this view is too unrealistic, and fails to do justice to the practice of scientific explanations. Oftentimes, we accept a scientific theory that include novel scientific statements as true or justified just because we think the theory leads to best explanations. That is, whether a scientific statement has explanatory power (or unifying power, according to Kitcher) is sometimes prior to the question of whether this scientific statement is reasonably believed to be true. So, it's too demanding to require that every statement used to unify K must be already accepted as true.

Moreover, for our purposes here, we don't have to require that the premises of K be already reasonably accepted as true. In the discussion of moral explanation, both the moral skeptic and the defender of moral explanation believe that only after we vindicated the explanatory power of moral facts can we justify moral beliefs—the question of whether moral statements have explanatory power is epistemically prior to the question of whether moral statements are reasonably accepted as true.

A Modest Version of Unificationism

The version of unificationism I will propose is more modest, and hopefully more plausible, than Kitcher's theory. Kitcher's unificationism has two distinctive features: it is both global and hard unificationism; in contrast, I advocate a local and soft version of unificationism. Let's consider the distinction between global and local unificationism first. Global unificationism is the view that an

explanatory pattern plays a genuine unifying role if it, together with other explanatory patterns, best unifies the whole body of human knowledge (or all reasonably accepted beliefs). In contrast, local unificationism asserts that an explanatory pattern plays a genuine unifying role if it provides a more unified account of a certain kind of phenomena than all other explanatory patterns.

Global unificationism is problematic for two reasons. First, global unificationism is either implausible or useless, because it (unrealistically) requires that we have the ability to compare competing systemizations within the whole body of knowledge in order to choose genuine unifying explanations. Second, global unification seems irrelevant or unnecessary for explanatory understanding; local unification is probably a better candidate. For example, suppose that biology plays a more unifying role in explaining biological phenomena than any other disciplines. Then it would be fair to say that biology contributes a distinctive kind of explanatory information and advances our understanding of biological phenomena. Why would it matter whether biological explanation is included in the set of explanatory patterns that maximally unifies the whole body of human knowledge? As long as I can show that moral properties play a more unifying role in explaining a certain kind of phenomena than descriptive properties, the explanatory efficacy of the moral could be vindicated.

Let us also distinguish between hard and soft unificationism. In my terminology, hard unificationism is the view that unification is the sole explanatory virtue and stands as the necessary and sufficient condition for explanatory efficacy—in other words, an explanatory pattern provides legitimate or valid explanations if and only if the explanatory pattern plays a genuine

unifying role. In contrast, soft unificationism is the view that unification is only a sufficient, but not necessary, condition for explanatory efficacy—that is to say, an explanatory pattern provides legitimate explanations if it plays a genuine unifying role.

I advocate soft unificationism (instead of hard unificationism) for two reasons. First, I believe that a pluralistic account of explanatory virtues is more plausible than a monistic account. Explanatory pluralism is the view that there are multiple and incomparable explanatory virtues, such as unification, simplicity and depth. Higher-order and lower-order explanations often exhibit different explanatory virtues: a lower-order explanation provides more details at a deeper metaphysical level about why the phenomenon occurs, whereas a higher-order explanation exhibits a more unifying feature in explaining the occurrence of the same phenomenon. According to explanatory pluralism, both explanations can be valid explanations and need not compete with each other—neither is better than the other. As Sober puts it:

> Higher-level sciences often provide more general explanations than the ones provided by lower-level sciences of the same phenomena… Generality is one virtue that an explanation can have, but a distinct and competing virtue is depth, and it is on this dimension that lower-level explanations often score better than higher-level explanations. The reductionist claim that lower-level explanations are always better and the antireductionist claim that they are always worse are both mistaken (Sober 1999).

From my point of view, I really doubt that we can compare, for example, unification and depth in a hierarchical way, just as the value pluralist doubts that happiness and wisdom, for instance, can be comparable in a similar way. While I won't be arguing for this intuition here, I would note that other philosophers

have given a convincing defense of it (see Railton 1981; Jackson and Pettit 1992; Sober 1999).

Second, for the purpose of saving moral explanation, I don't need to hold a hard version of unificationism. My vindication only requires the truth of soft unificationism. As long as I can establish that moral facts play a distinctive unifying role in moral explanations, I would show that some explanations that appeal to moral facts are no less valid than any other explanations that assume no moral facts. Then I would vindicate the explanatory efficacy of moral facts in this way.

Unification by Multiple Realization

Certainly, an important way of unifying phenomena is by *reduction*. For example, physics provides a unified account of different phenomena by reducing macrophysical facts or properties to microphysical facts or properties. Or, we can consider Michael Friedman's example of the kinetic theory of gases. This theory explains phenomena involving the behavior of gases, such as the fact that gases approximately obey the Boyle-Charles law and the fact that they obey Graham's law of diffusion, by reference to the behavior of the molecules which obey the laws of mechanics. So the kinetic theory effects a significant unification: where we once had some independent brute facts that gases obey the Boyle-Charles law and that they obey Graham's law, we now have only one fact that molecules obey the laws of mechanics (Friedman 1974, p. 14).

What I wish to stress here, however, is that although reduction is a significant form of unification, it is not the only way. If some higher-order

properties are multiply realized by lower-order properties, the higher-order properties could play a more unifying role in explanations than their base properties do. Consider an example. Suppose that different creatures' pain is realized by different neural or physical states. For instance, dogs' pain is realized by physical state P_1, bats' pain by P_2, humans' pain by P_3, and Martians' pain by P_4.... And *suppose* that when these creatures are in pain, they all tend to *wince*.[28] Imagine a dog which is in pain and hence winces. If someone asks: "Why does the dog wince?", we can explain the event by appeal to either the mental property, pain, or the base neural property P_1. Then we have two proposed argument patterns, A1 and A2:

(1)	Individuals of type C tend to do W when they experience pain;
(2)	*m*, an individual of type C, experiences pain;
(C)	Therefore, *m* does W.

(1*)	Individuals of type D tend to do W when they are in physical state P_1;
(2*)	*m*, an individual of type D, is in physical state P_1;
(C*)	Therefore, *m* does W.

Here D is a subset of C. The first argument pattern (A1) thus seems to play a more unifying role in explaining creatures' wincing behavior than the second argument pattern (A2). For we can use the first pattern to explain why a bat (say) which is not in P_1 but P_2 also winces, but we cannot use the second pattern to do so. Now we can clearly see how explanations involving multiply realizable properties exhibit a unifying feature.

As I have discussed above, multiply realizable properties, such as pain, can play a more unifying role in explanations than their realizing properties.

[28] By 'wincing', I mean a shrinking body movement.

Therefore, higher-order disciplines can still exhibit irreplaceable explanatory power, which wouldn't be replaced or excluded by the explanatory power of lower-order disciplines. As Peter Railton puts it:

> Psychologists, sociologists, and biologists have formulated worthwhile theories using the distinctive categories of their disciplines. Such theories contain generalizations that may not be strict or exceptionless, but that do illuminate functional connections, causal dependencies, and other relations at a particular level of description of the phenomena. These relations afford explanatory insights that would not be evident at the (subjacent) level of a purely physical description of events (Railton 1998).

The same is true of moral explanation. Consider a moral property, say, injustice. It seems that what makes a society unjust can be very divergent, such as racial discrimination, the lack of political rights, and arbitrary rulings. And suppose that different forms of injustice would all *tend* to cause a revolt. Consider two societies. Racial discrimination prevails in one society, and the other society lacks political rights. Suppose that a revolt occurred in each society. If someone asks: "Why is there a revolt in each society?", we can give a unified explanation by saying that they are both unjust. But if we don't use the moral property, injustice, but instead appeal to the realizing properties of injustice to do the explanation, then we will appeal to two less unifying explanatory patterns that involve racial discrimination and lack of political rights respectively. Consider two proposed argument patterns, A3 and A4:

(a)	Societies of type I tend to involve R when they are unjust;
(b)	*t*, a society of type I, is unjust;
(C)	Therefore, *t* involves R.

(a*)	Societies of type L tend to involve R when they lack political rights;
(b*)	*t*, a society of type L, lacks political rights;
(C*)	Therefore, *t* involves R.

Here L is a subset of I. The first argument pattern (A3) thus seems to play a more unifying role in explaining revolutions than the second argument pattern (A4). For we can use the first pattern to explain why another unjust society which is in a different descriptive state also involves revolution, but we cannot use the second pattern to do so.

Someone might reply that although the unifying power of a multiply realizable property goes beyond the unifying power of any of its realizers, the disjunction of its realizers would exhibit the same range of unifying power. However, this disjunctive move is problematic for two reasons. First, good explanations are those which appeal to genuine properties. But many philosophers worry that merely disjunctive properties are themselves not genuine properties.

Second, no matter whether a merely disjunctive property is a genuine property, it seems that explanations that appeal to such merely disjunctive properties are unsatisfactory or spurious. A multiply realizable property like pain helps capture 'patterns' that the disjunctive property fails to capture—a multiply realizable property provides some significant information that the disjunctive property does not. For example, when we observe that a dog that is in state P_1, a bat that is in state P_2, a human that is in state P_3, and a Martian that is in state P_4... all tend to wince, we must ask: what is common to P_1, P_2 and P_3... such that different creatures which are in different physical states behave in the same way? To answer that these physical states are all disjuncts of the disjunctive property P_1 v P_2 v P_3... is not interesting, on the one hand. On the other hand, to reply that all of them are realizers of the mental state, pain (so the creatures will tend to wince)

eventually suggests a psychological explanation rather than a purely physical explanation (see Fodor 1974, 1997; Pereboom and Kornblith 1991).

In a nutshell, moral properties could play a distinctive unifying role in explanations, and the argument patterns that appeal to moral properties can provide legitimate explanations and advance our overall understanding. The explanatory role of moral properties thus goes beyond the explanatory role of their descriptive realizers.

Functional Unification and Moral Explanation

Different kinds of properties (categorical properties, dispositional properties, or functional properties, etc.) may play different unifying roles in explanations. Here I want to say something about functional unification, i.e. the way of unification by functional properties. In what follows, I attempt to suggest that moral properties understood as functional properties can play a distinctive unifying role in moral explanations.

In recent decades, quite a few naturalistic moral realists, such as Richard Boyd, David Brink, Frank Jackson and Peter Railton, introduce functionalism into ethics for characterizing the nature of moral properties (Brink 1984; Boyd 1988; Railton 1993; Jackson 1998). For example, David Brink writes:

> [T]he moral realist might claim that moral properties are functional properties. He might claim that what is essential to moral properties is the causal role which they play in the characteristic activities of human organisms. In particular, the realist might claim that moral properties are those which bear upon the maintenance and flourishing of human organisms. Maintenance and flourishing presumably consist in necessary conditions for survival, other needs associated with basic well-being, wants of various sorts, and distinctively human capacities. People, actions, policies, states of affairs, etc. will bear good-making properties just insofar

> as they contribute to the satisfaction of these needs, wants, and capacities… [and] will bear bad-making properties just insofar as they fail to promote or interfere with the satisfaction of these needs, wants, and capacities (Brink 1984, pp. 121-2).

As I will discuss in more detail in Chapter 5, I propose a particular version of moral functionalism, which understands moral properties as second-order, empirical, and holistic functional properties. But here I will just focus on moral functionalism in general. My functionalist approach to the problem of moral explanation is inspired by contemporary philosophy of mind. According to psycho-functionalism that is influential in this area, mental properties should be identified in terms of certain functional roles, and hence they are multiply realized by, and hence irreducible to, their physical realizers (e.g. physiological properties and neural properties). Psycho-functionalism shows further that irreducible mental properties are still natural, empirically accessible properties, and play an *irreplaceable* role in the explanations of human actions and social events—the irreducibility, the naturalness and the explanatory efficacy of the mind can happily live together. It is important to note that these accounts demonstrate how irreducible functional properties in general can nonetheless have explanatory power.

In particular, moral functionalism supports a unificationist vindication of moral explanation in three ways. *First*, since it is a general consensus among philosophers that there are multiple, even indefinite and infinite, base properties that can occupy a certain functional role, moral functionalism explains why moral properties are multiply realized by purely descriptive properties in the way that functionalism about the mind explains the multiple realizability of mental

properties. Actions, characters, etc. that have different descriptive properties can all be morally good (say). What do these descriptive properties have in common? A plausible answer is that these descriptive properties occupy the same functional role that defines goodness. Characterizing moral kinds as functional kinds seems to be at exactly the right level of abstraction to handle multiple realizability. Therefore, moral functionalism grounds a unificationist vindication of moral explanation by accounting for multiple realizability.

Second, functional unification reveals *intelligible* explanatory connections (rather than bare causal regularities) and hence advances our understanding. For example, given that the functional role of pain essentially involves aversiveness, we can clearly see why a creature tends to wince when it is in pain no matter which particular neural state realizes pain. Being in pain thus contributes a more intelligible explanation of wincing than the particular neural state. By the same token, given that the functional role of injustice essentially involves, say, discounting the interests of certain groups, we can clearly say why a society tends to be in revolt when it is unjust no matter which particular social/political state realizes injustice. So injustice is more explanatorily relevant to the occurrence of the revolt than the particular social/political state. To see *only* the regular causal connection between the particular descriptive state and revolt (or the particular neural state and wincing) would be to miss the main explanatory point: *injustice* foments revolt (or *pain* makes one wince).

Third, functional unification typically provides us with certain *modal* information about why the explanandum phenomenon occurs.[29] For example, suppose that the property of having such and such neurons firing in the anterior cortex (call this property 'P') can play the functional role of pain on a particular occasion. Of course, the 'pain role' can be implemented in numerous different ways; neuroscience suggests that other similar properties, such as a property P' that is exactly like P except in which a particular firing neuron is different, can also play the functional role of pain. In the counterfactual circumstances where pain is realized by P' rather than P, a person would still have winced. Similarly, the functional role of injustice can be implemented by different descriptive properties. So, even if a society were in another similar but different social/political state that still implements the functional role of injustice, there would have been also a revolt in the society. That is, a functional explanation helps to show how the actual world runs on patterns found in a variety of possible worlds—functional unification reveals the 'what-if-things-had-been-different' aspect of explanation (Woodward 2003). Such explanations that appeal to a functional property unify actual cases with counterfactual cases by offering some modal information about the occurrence of the effect.

[29] To be more accurate, functional explanations, as compared with explanations that appeal to first-order realizers, provide *modally comparative* information. As Jackson and Pettit indicate, there are two grossly different sorts of explanatory information. One sort of information helps us to *differentiate* the actual world from other ways the world might have been (i.e. other possible worlds). The other sort of information helps us to *relate* the actual world to other possible worlds. The first sort of information is modally contrastive, and the second modally comparative (see Jackson and Pettit 1992).

Dissolving a Putative Disanalogy

I have already outlined a unificationist vindication of moral explanation. In the rest of the chapter, I want to discuss a putative objection to my proposal. The moral anti-explanationist may contend that even though moral properties play a more unifying role in explaining some phenomena than their descriptive base properties, the unifying role of moral properties could be replaced by that of people's psychological representations of moral properties. In what follows, I will attempt to argue that this objection does not in the end raise serious challenges to the explanatory efficacy of the moral.

Some philosophers claim that putative moral explanations are always mediated by people's moral beliefs, attitudes or sensibilities etc. (henceforth I will speak of 'moral beliefs' for short). Thus, the anti-explanationist argues, if these purported moral explanations are at all informative, it is only because such accounts tell us indirectly about people's moral beliefs. For example, in appealing to injustice to explain the revolt, says the anti-explanationist, what actually explains is just people's belief about injustice. Here the moral term 'injustice' appears to occur only obliquely: it occurs only in giving the content of people's moral convictions, not in describing some putative state of the world. Blackburn writes:

> Someone citing injustice as the cause of revolution might be adverting to the population's *perception* of injustice, or belief that they are victims of injustice. And there is nothing surprising in this having effects. This is a plausible construction of the revolutions case: at least, I cannot readily think of mechanisms whereby injustice brings about revolutions except through the population's awareness of it (Blackburn 1991, pp. 11-2).

Admittedly, the explanatory role of other higher-order properties is sometimes also mediated by people's beliefs about those properties. For example, I'm in pain, and I believe that the best way to eliminate my pain is to see a doctor. Hence I walk to the hospital. In this case, the explanation of my walking is mediated by my belief about pain. But it is important to note that the explanatory role of pain is not *always* mediated by that of pain-beliefs. For instance, many people tend to sweat when they are in pain. The explanation of their sweating doesn't seem to be mediated by their beliefs about pain.

Therefore, since moral explanations are always mediated by moral beliefs, says the anti-explanationist, this gives us a good reason to believe that the putative explanatory power of moral properties is fully excluded by that of moral beliefs. This time what competes with the explanatory power of moral properties is not their descriptive realizers, but the psychological states involving the attribution of moral properties; putative moral explanations threaten to be reducible to psychological explanations of some sort. For example, regardless of whether the explanatory power of injustice is excluded by that of the realizers of injustice such as lack of political rights, it might be still argued that the explanatory role of injustice would be excluded by the explanatory role of people's beliefs about injustice. Even though injustice plays a more unifying role than any of its realizers in explaining a revolt, it *seems* that injustice does no better job than people's beliefs about injustice. Injustice-beliefs appear to play a unifying role of the same range in explaining the revolt: when people observe any of the base properties of injustice, they judge, consciously or unconsciously, that

that is unjust and then act accordingly. How, then, can the moral explanationist conjure up unificationism to vindicate the explanatory efficacy of injustice?

A straightforward way of objecting to the second disanalogy is by pointing out that moral explanations are NOT always mediated by people's moral beliefs. Peter Railton famously advocates this approach. He writes:

> Explanations involving assessments of social rationality cannot be replaced by explanations involving *beliefs about* what would be morally right. For example, discontent may arise because a society departs from social rationality, but not as a result of a belief that this is the case. (Railton 1986, p. 192)

> A social order's injustice may produce alienation and dysfunction well before any articulated sense of its injustice surfaces within the population. A parent's dishonesty may undermine the solidity of a child's sense of self long before any glimmer of a moral critique of the parent emerges in the child's mind… (Railton 1995, p. 273)

Nonetheless, the anti-explanationist might reply that the crucial question here is not whether moral explanations are always mediated by moral beliefs; rather it is whether moral properties can still be explanatorily potent even in cases which are mediated by moral beliefs. Other sorts of higher-order explanations of human behavior are sometimes also mediated by the role of beliefs or similar intentional states, but it seems that higher-order properties can still have explanatory power in these 'doxastic' cases (i.e., cases where the explanations are mediated by beliefs or similar intentional states). In the earlier example, my being in pain seems to truly explain my walking to the hospital, even though this explanation is certainly mediated by my belief about pain. If moral properties were explanatorily impotent in all doxastic cases, there would still be a disanalogy between moral explanations and other sorts of higher-order explanations (Zangwill 2005). What's more important, if the explanatory role of moral properties were excluded

by that of moral beliefs in doxastic cases, this failure would cast doubt upon the explanatory power of the moral in general—the putative explanatory role of the moral in non-doxastic cases would be also suspicious.

I will offer two unificationist responses to this challenge: one is short, and the other is long. First, since we already agree that some putative moral explanations are not mediated by moral beliefs, the unificationist strategy can be introduced to defend moral explanation in general. Because an argument pattern that appeals to a moral property M can derive phenomena in both doxastic contexts and non-doxastic contexts (contexts which are not mediated by M-belief), the argument pattern that appeals to M can unify phenomena better than an argument pattern that appeals to M-belief. Thus, the explanatory role of M is not excluded by that of M-belief.

Second, the moral explanationist can argue that moral properties play a legitimate role in explaining our moral beliefs—moral properties might somehow figure in the *etiology* of our moral beliefs. In the example of pain, although the explanation is mediated by my belief about pain, we have the strong intuition that my being in pain still plays some explanatory role. The reason is probably that pain is explanatorily relevant to my belief about pain—it is my being in pain that helps explain why I have the belief that I am in pain. By the same token, if we can show that moral properties are relevant to explaining the occurrence of our moral beliefs, then we can vindicate the explanatory efficacy of the moral even in cases which are mediated by moral beliefs.

Let's consider the phenomenon of normative change. Human history has seen a great change in people's moral beliefs. Take beliefs about justice as an

example. People's conceptions of justice in the pre-modern times were very different from contemporary views. It was not unusual that pre-modern societies regarded, say, slavery, dictatorship, brutal punishment, criminalizing homosexuals, and religious intolerance as just (or at least morally permissible). In contrast, modern societies generally hold very different moral beliefs about justice, according to which a just society should, for example, separate state from church, respect human rights, endorse constitutional democracy, treat its citizens as equals, and so forth.

It is important to note that this change in justice-beliefs is not a local phenomenon happening in a few countries, but rather a common trend across almost the whole world. It is hard to believe that the convergence of people's beliefs about justice, at least at a fairly general level, is just a mere coincidence. Rather, this phenomenon needs an intelligible explanation. Someone might suggest that we can appeal to a so-called 'social selection' account to explain why there is a common trend converging on modern beliefs about justice. On this account, whereas the descriptive social/political conditions people *previously* identified as just tended to prevent the stability and flourishing of a society, the descriptive social/political conditions we *now* identify as just help promote stability and flourishing. As a result of this social selection, modern beliefs about justice have persisted and expanded over the world. In short, the social selection account appeals to these descriptive conditions or properties to explain the stability and flourishing of societies, and hence to explain the change in justice-beliefs. The moral anti-explanationist could argue that we can explain the

widespread presence of modern beliefs about justice without any need to posit a role for justice itself.

However, I want to suggest that if we look at the issue more closely, we would find that the unificationist strategy can be used to vindicate the explanatory role of justice in explaining stability and flourishing (and hence explaining the change in justice-beliefs). My basic point is this. First, the moral explanationist can offer a normative ethical theory of justice, according to which justice tends to promote stability and flourishing. For example, a just society typically treats its members as equals, distributes the goods in a fair way, and tolerates permissible ways of life, etc. Thus, just societies tend to be stable and flourishing. Second, the normative theory of justice could show that justice is multiply realizable by different descriptive properties in different societies, actual or counterfactual. For instance, just societies of different sorts might emphasize different '*currencies*' (or metrics) of distribution, such as capabilities, happiness, and primary goods. Even societies that have the same distribution metric could adopt different distribution *patterns*, which still realize justice.

Call those social/political realizers of justice N_1, N_2, N_3…. If we use justice to explain the stability and flourishing of modern societies and hence the change in justice-beliefs, then we could provide a distinctive unifying explanation. But if we solely use these descriptive properties to do the work, then we will either provide less unifying explanations that appeal to individual justice-realizers N_1 or N_2 or N_3…, or give an awkward disjunctive 'explanation' that appeals to the infinite or indefinite disjunctive property 'N_1 v N_2 v N_3…'. Insofar as some society instantiates justice, it tends to promote stability and flourishing, no matter which

particular social/political feature realizes justice. Hence justice contributes a distinctive kind of explanatory information and advances our overall understanding. Since justice is explanatorily relevant to our beliefs about justice, the explanatory role of justice won't be excluded by that of people's beliefs about justice. To generalize, we could vindicate the explanatory efficacy of moral properties in doxastic contexts.

Someone might ask: "How do you *know* that such and such social/political conditions are just?" "Are you really *justified* in believing that?" My quick reply is that the question of whether moral beliefs are justified cannot be answered prior to the question of whether putative moral facts are explanatorily efficacious, just as the question of whether our beliefs about microphysical entities (say) are justified cannot be answered prior to the question of whether supposed microphysical facts have explanatory power. It would be unfair to ask the explanationist to give a complete justification of moral beliefs in the first place—moral beliefs can be justified only after moral explanations are vindicated.

If someone already assumes that it's unreasonable to believe in moral facts, then it would trivially follow that it is unreasonable to believe that those facts have explanatory power. But this is not an interesting objection to moral explanation. Moreover, many anti-explanationists attempt to deny moral knowledge precisely by rejecting moral explanation. If they were then to argue against moral explanation by assuming that we don't have moral knowledge (or justified moral beliefs), they would be doing something very much like begging the question.

Conclusion

As I have argued in this paper, moral properties can have genuine explanatory power in virtue of their distinctive unifying role (even if they are irreducible to descriptive properties). But I wish to stress that I don't believe that higher-level explanations are better than lower-level explanations in any 'absolute' sense. According to the explanatory pluralism I accept, there are incomparable explanatory virtues such as unification and depth, and different explanations that exhibit different virtues don't outweigh each other (for example, I can agree that the 'social selection' account and the moral explanationist account I propose explain the change in justice-beliefs *equally well*). Unifying higher-level explanations and deep lower-level explanations should cooperate and jointly contribute to our overall understanding. Then the explanatory efficacy of the moral will be appreciated in such an 'ecumenical' world.

Chapter 5

A Non-reductive Naturalist Version of Moral Functionalism

Introduction

In the earlier chapters, I have mentioned moral functionalism on a number of occasions. In the last chapter of my dissertation, I want to defend a non-reductive version of moral functionalism and give an account of what kind of irreducible natural properties moral properties may probably be.

Since the last part of 20th century, functionalism has stood as the dominant theory of mind. Functionalism in the philosophy of mind is the doctrine that what makes something a mental state of a particular type doesn't depend on its intrinsic or qualitative features, but rather solely on the way it functions, or the role it plays.[30] More precisely, functionalism takes mental states as functional states that are specified in terms of their roles as causal intermediaries between inputs and outputs (including sensory stimulations, behaviors and other mental states). Let me take pain as an example. Pain is typically caused by bodily injury or disturbance (e.g. cut, burn and pressure).

[30] Functional properties, by definition, are those that are specified in terms of functional roles. But it is compatible with the definition of functional properties that the characterization of functional roles may involve some non-functional properties or entities. For example, suppose that pain is defined in terms of a certain functional role. And suppose that the characterization of 'pain' role includes a causal statement that "pain is caused by C-fiber firing". Although C-fiber is probably not a functional property, this is coherent with pain's being a functional property.

Pain is aversive and typically evokes a negative attitude or desire that the sensation should cease. Pain is generally distracting, and intense pain even has the tendency to cause wincing or groaning.... We can call such inputs and outputs of pain the 'functional role of pain', which is used to characterize what pain is. Any psychological state-token that satisfies the 'pain' role should be thus considered as an instance of pain.

Recently, inspired by functionalism in the philosophy of mind, some naturalistic moral realists, such as Richard Boyd, David Brink, Frank Jackson and Peter Railton, introduce functionalism into ethics for characterizing the nature of moral properties (Brink 1984; Boyd 1988; Railton 1993; Jackson 1998). Just like psychological functionalism, moral functionalism understands moral properties as functional properties of some sort, which are specified solely in terms of their functional roles.[31] Take rightness as an example. Rightness typically bears on the promotion of people's well-being. Rightness is intrinsically motivating. When one judges an action to be right, she normally has a prima facie motivation to perform this action. Also, rightness plays a normative role in guiding human actions. An action's being right provides people with an important, even overridingly so, normative reason to do it.... Thus we can appeal to this functional role to define what rightness is.

As I have discussed in earlier chapters, moral functionalism is attractive for several reasons. First, a functional account of moral properties can explain, or

[31] Although a non-naturalistic or supernaturalistic version of moral functionalism is conceptually possible, almost all moral functionalists are ethical naturalists. Thus, in my dissertation, I will always use the notion of "moral functionalism" in the sense of naturalistic moral functionalism.

cohere with, other widely-accepted theses, such as moral-natural supervenience[32] and the multiple realizability of moral properties.[33] Second, as I argued in Chapter 3, if moral properties are functional properties, the causal power of moral properties would be preserved. Or, at least, moral causation would be no more mysterious than other kinds of functional causation, which are usually assumed in special sciences. Third, as I mentioned in Chapter 4, moral functionalism can help vindicate the explanatory efficacy of moral properties, since functional explanation, which is also generally adopted in many scientific disciplines, can play a distinctive unifying role.[34]

In particular, I will propose a version of moral functionalism. *First*, my version of moral functionalism is a kind of second-order moral functionalism rather than first-order moral functionalism. According to second-order moral functionalism, moral properties are identified as the second-order properties of having some properties that play certain functional roles, but not as the first-order occupants or realizers that actually play those roles. *Second*, my moral functionalism is a form of synthetic moral functionalism, according to which the functional roles of moral properties can only be specified by appeal to empirical investigation, as opposed to analytic moral functionalism, the view that moral

---

[32] If moral properties are functional properties, then it seems to follow that moral properties supervene upon natural properties, since it is reasonable to assume that natural properties are the realizers of functional properties.

[33] Since it is very unlikely that there is only one base property that can realize a certain functional role, functional properties are standardly regarded as being multiply realized by their base properties. So moral functionalism can do justice to the multiple realizability of moral properties.

[34] I know that it is controversial whether functional properties have causal power or explanatory power. For example, some philosophers claim that the causal power of second-order, functional properties would be excluded by their first-order realizers. Also, some philosophers complain that functional explanations are derivative, at best, or uninformative, at worst. However, as I have argued in Chapter III, IV, functional properties as second-order properties can figure in legitimate causal explanations and play a unifying explanatory role.

properties can be identified as certain functional properties solely by conceptual analysis. *Third*, my version of moral functionalism is a holistic version of moral functionalism, according to which the functional roles of moral properties cannot be captured without reference to other moral or normative properties. Since my version of moral functionalism is *second-order*, *synthetic*, and *holistic*, I will call it 'SSH moral functionalism' for short. It is important to note that SSH moral functionalism is a form of non-reductive ethical naturalism. On the one hand, a synthetic, *a posteriori* version of moral functionalism is a sort of ethical *naturalism* in the sense that the nature of ethical properties is subject to empirical investigation. On the other hand, as I will discuss later, a second-order, holistic version of moral functionalism is a kind of ethical *anti-reductionism*. In this chapter, I will attempt to argue that SSH moral functionalism is more tenable than alternative versions of moral functionalism. Moreover, I will consider some criticisms of moral functionalism, such as the 'moral twin earth' challenge, and then attempt to show that those objections are not serious threats to SSH moral functionalism.

First-order vs. Second-order Moral Functionalism

As moral functionalism is substantively influenced by the philosophy of mind in which functionalism is well developed and widely discussed, it might be very helpful to consider functionalism about the mind first. Let me take 'pain' as an example again. According to functionalism in the philosophy of mind, for an organism to be in pain is for it to be in some internal state that is typically caused by bodily injury and that typically causes wincing or moaning or certain negative

attitudes. However, there is a further question to be clarified, which is not settled by the above account: Is pain a first-order property, i.e. the property or state that actually plays the 'pain role', or a second-order property, i.e. the property of having some property, or being in some state, that plays such and such role?[35] Some philosophers, such as Armstrong and Lewis, seem to adopt the first option, whereas other philosophers, e.g. Putnam, Fodor and Kim, take mental properties as second-order properties (see Fodor 1965; Armstrong 1968; Lewis 1972; Putnam 1975; Kim 1978). I call the first view 'first-order functionalism' and the second view 'second-order functionalism'.[36]

Kim gives a definition of the notion of 'second-order property'. Let B be a set of first-order properties. We then have this:

> F is a second-order property over set B of first-order properties if and only if F is the property of having some property P in B such that D(P), where D specifies a condition on members of B (Kim 1998, p. 20).

For example, if B is the set of colors, then the property of having a primary color can be considered as a second-order property: having a property P in B such that P = red or P = blue or P = green. Second-order properties are second-order in this sense that they are generated by *quantification* over the first-order properties. We can call the first-order properties satisfying condition D the realizers of second-order property F, just as being red, being blue, and being green are the three realizers of having a primary color (Kim 1998). Similarly, Kim regards functional properties as second-order properties whose specification involves

---

[35] The property of being *in* some state that plays such and such role is a *relational* property. But as I will mention later on, the second-order functionalist may regard pain (or any other mental property) as a non-relational, second-order property.

[36] First-order and second-order functionalism are also called 'the functional specification theory' and 'the functional state identity theory', respectively. See Block 1980a.

functional roles of first-order properties. An example of a functional property is solubility: a substance has this property just in case it has some physical property or other P that causes this substance to dissolve in water—in this case P is the realizer of the property of solubility.

In what follows, I will argue against first-order functionalism and hence show that mental properties should be understood as second-order properties rather than first-order properties. Suppose that the functional role of pain (say) is played by C-fiber stimulation. Then, according to first-order functionalism, pain *is identical with* C-fiber stimulation. So, on this view, if a cat and I are in pain, we must be in the same physical state. However, this view goes against the widely-accepted thesis of multiple realizability. Given the diversity of organisms, it is very unlikely that all creatures capable of consciousness or experience must have their physical states in common. Moreover, even if pain is singly realized in the *actual* world, it is quite conceivable that pain could be realized by different physical states in other possible worlds—but (property) identity entails *necessary* coextension. So, it is safe to say that the same mental state is multiply realized by different physical states, just as sugar and salt which have different micro-physical structure share some common higher-order properties, such as 'water-solubility' (see Putnam 1975).

Someone might deny that there are any species-independent (or population-independent) mental states. For example, according to early Lewis, there really is no such state as the state of pain *simpliciter*; rather, there is only human pain, bat pain, Martian pain, and so on (Lewis 1969). On this suggestion, multiply realizability doesn't threaten the identity of mental states and physical

states. But I don't think that this suggestion is satisfactory. Admittedly, we can regard almost any two things, even two tokens of the same type, as different in virtue of some of their qualities, intrinsic or extrinsic. Of course, we also can consider pain-in-humans and pain-in-cats to be different, but, why do we call them 'pain'? What is common to the pains of humans and cats in virtue of which they are pains, i.e. in virtue of some property only pain states have?[37] According to our common-sense classification of mental states, we regard both pain-in-humans and pain-in-cats as pain whatever the criterion of classification is. This indicates that pain in different species probably have some distinctive features common to all pain states. I suggest that we should respect our common-sense classifications of mental states, unless we have good reasons not to do so.

In order to respond to the challenge of multiple realization and at the same time respect our common-sense classifications of mental states, the first-order functionalist might suggest that 'pain', for example, be a non-rigid designator, which behaves roughly as an indexical and picks out different physical states in different species (or even different individuals) that play the same functional role. Thus,

> pain = def. the state that plays functional role R in relevant population P at time *t*.

On this view, pain states do have in common some property that is essential to all pain states, that is, the property of playing role R in appropriate population at a certain time.

[37] Certainly, any two things have infinite properties in common, but what the first-order functionalist need find is a property that is essential to all pain states.

But wait a minute! If the property of playing role R is essential to all pain states, then it would follow that when creatures are in pain, what is common to them is just the property of being in some state that plays role R, the context-independent, second-order property which second-order functionalism (normally) identifies as pain. Since the first-order functionalist concedes that this second-order property is essential to pain, isn't she actually committed to second-order functionalism, according to which pain is a second-order property?

The first-order functionalist could reply that although this second-order property is essential to creatures' being in pain, it doesn't follow that pain itself is a second-order property. She might conjure up her last weapon by distinguishing 'pain' from 'being in pain'. Whereas 'being in pain' is a rigid designator that refers to a second-order relational property, i.e. the property of being in some state that plays the functional role, 'pain', on this view, rather picks out the first-order physical states in relevant populations that play the functional role. And it may be argued that pain, rather than being in pain, is the mental state under consideration (Tye 1983). This distinction between pain and being in pain appears to be supported by features of English language. Consider again the case of pain. The grammatical structure of the predicate 'is in pain' naturally suggests that it expresses a relational property, that of bearing the 'in' relation to the state of pain. Hence it seems that the predicate 'is in pain' cannot express the very thing denoted by the term 'pain', since the latter is simply the state of pain itself. Once this distinction is drawn, it is tempting to think that similar distinctions should be drawn between other mental properties and states, for example, the

distinction between anger and having anger, the distinction between fear and having fear, and so on.

But I doubt that this response is plausible. I have two further replies. First, I worry that this suggestion may be misled by English grammar. As some philosophers point out, when we say "I'm in pain", what we actually mean is just "I hurt", or "I'm pained". 'Pain' seems to denote the property of hurting or the property of being pained; the linguistic distinction between pain and being in pain doesn't correspond to a genuine conceptual distinction. This point can be better captured if we have a look at other languages, which don't support such a distinction. For example, in Chinese, there is no such linguistic distinction between 'pain' and 'being in pain', or between 'fear' and 'having fear'. Chinese speakers simply use non-relational phrases to express the meaning of 'being in pain' or 'having fear' in English: they use "我疼" ("I am pained" or "I hurt") to express "I'm in pain", and use "我害怕" ("I fear") to express "I have fear". Thus, I doubt that the linguistic distinction in English between 'being in pain' and 'pain' is good evidence that there really exists a corresponding conceptual distinction between two properties.

Second, even if there is such a conceptual distinction between pain and being in pain, how does this help to show that pain is a first-order property? Even if pain is not a *relational*, second-order property, i.e. the property of being in some state that plays role R, it can be a *non-relational*, second-order property, i.e. the property of being a state that plays role R. The latter is definitely a second-order property of the first-order physical states. Although many second-order

functionalists identify pain with the relational property, I see no reason why we couldn't understand pain as the non-relational one. Moreover, whether or not pain is the property of being a state that plays role R, it seems to me that the second-order, non-relational property also forms a psychological kind.[38] And since the second-order, non-relational property is not reducible to its first-order physical states, first-order functionalism is probably false. For it is the view that all mental states can be functionalized and hence reduced to first-order physical states.

Therefore, I conclude that second-order functionalism about the mind is more tenable than its first-order competitor. The same is true of moral functionalism. For similar reasons, we should understand a moral property as a second-order functional property, i.e. the property of having some property that plays such and such functional role. Moreover, there are additional, independent reasons to adopt a second-order version of *moral* functionalism. First, it is more natural to understand moral properties as second-order properties rather than their first-order realizers, since English language doesn't suggest such a distinction, for example, between goodness and being in goodness or having goodness. Second, we have a robust intuition that moral terms are rigid designators.[39] For example, the term 'wrong' is supposed to pick out the same property in every possible world. But if wrongness were a first-order realizing property, the term 'wrong' would pick out different properties in different possible worlds, given the fact of multiple realization. Since second-order

[38] I suspect that the second-order, relational property (that is, the property of being in pain) is also a psychological property, even if being in pain is distinguished from pain itself.
[39] I use the term 'rigid designator' in a broad, loose sense, according to which a rigid designator could be a property term, not just a substance term or a proper name.

functional properties are denoted rigidly but first-order properties are not, moral properties are better understood as second-order functional properties of some kind. Moreover, as I will discuss later, second-order functionalism (rather than first-order functionalism) can help meet the Moral Twin Earth challenge by solving the problem of indeterminacy.

Analytic vs. Synthetic Moral Functionalism

As I discussed in the last section, we should adopt a version of second-order functionalism both in moral and psychological areas. In what follows, I will use 'functional properties' in the sense of second-order properties. But there is a further question: How do we come to know what functional properties mental properties (or moral properties) are?

It is worth mentioning a notorious distinction between two kinds of functionalism in the philosophy of mind: *a priori* functionalism (or analytic functionalism) and *a posteriori* functionalism (or synthetic functionalism).[40] According to analytic functionalism, mental states can be known as functional states with certain functional roles solely by appeal to the conceptual analysis of mental terms.[41] By contrast, synthetic functionalism claims that we must appeal to empirical investigation to discover what functional roles mental properties play (Block 1980a).

---

[40] We should notice that the analytic/synthetic distinction doesn't correspond to the *a priori*/*a posteriori* distinction. So, strictly speaking, *a posteriori* functionalism is conceptually different from synthetic functionalism. But here we can just ignore the difference.

[41] The notion of conceptual analysis is far from clear. Here I use 'conceptual analysis' in this way: if A is shown to be B after a (successful) conceptual analysis of 'A' and 'B", then the sentence "A is B" is analytically true or *a priori* true.

Similarly, moral functionalism can be also divided into two camps: analytic moral functionalism (e.g. Frank Jackson) and synthetic moral functionalism (such as Richard Boyd and Peter Railton). According to synthetic moral functionalism, the functional roles of moral properties can only be discovered by empirical investigation. According to analytic moral functionalism, we can know that moral properties are identical with certain functional properties solely by appeal to conceptual analysis.[42] Jackson and Pettit suggest that we appeal to conceptual truths or commonplaces to capture the functional role of a moral property, say, 'fairness':

> Such commonplaces are candidates for *a priori* truths: they are putatively such that anyone who knows how to use the term 'fair' is in a position to see that they hold… Those that are *a priori* true will give us the conditions under which fairness is instantiated (Jackson and Pettit 1995, p. 23).

I maintain that there are two reasons to prefer synthetic moral functionalism over analytic moral functionalism. First, I pretty much doubt that we can capture the 'real essences' of moral properties (here, the functional roles of moral properties) by conceptual analysis alone. The recent decades have witnessed the emergences of the so-called 'new wave moral realism'. According to the new wave moral realism, moral properties are natural kinds or quasi-natural kinds, which must be defined by *a posteriori* real essences rather than by purely conventional, *a priori* 'nominal essences'.[43] Only if we could categorize properties according to their hidden real essences, would systematic general knowledge of properties be possible. For this reason, we cannot know the functional roles of moral properties

[42] It is worth noticing that some analytic moral functionalists regard moral properties as first-order realizing properties rather than second-order properties (see Jackson and Pettit 1995). But conceptually there could be a second-order version of analytic moral functionalism.
[43] Natural kinds in a broad sense include not only natural substance-kinds, but also natural property-kinds. Here I use the term 'natural kind' in this broad way.

only by analyzing moral concepts; instead, we can only figure this out by empirically investigating the real essences of moral properties. Richard Boyd writes:

> According to 'naturalistic' approaches, the referent of a scientific (moral, theological…) term is defined by the properties necessary to underwrite a particular causal (or quasi-causal) role associated with the use of that term. On the conception I defend…, the relevant role is not fully specifiable *a priori,* nor is it an *a priori* question whether anything fulfills the role in question, or what properties underwrite its fulfillment when it is fulfilled. So naturalistic conceptions of reference share with psychofunctionalist conceptions of the referential semantics of mental terms a rejection of analytic definitions for the terms in question (Boyd 2003, p. 26).

I don't plan to discuss the first reason in more detail. In the rest of this section, I will focus on the second reason. I wish to stress that analytic moral functionalism, as a version of analytic ethical naturalism, also faces the challenge of Moore's open question argument. Moore thinks that moral terms cannot be analytically defined in terms of natural terms—moral terms cannot mean the same as any natural terms—since the question of whether an action or object described in terms of natural concepts was good or right was always open even to a competent speaker who grasps the meanings of the terms. For instance, a competent user of moral terms might believe that an action can promote an agent's happiness, but meanwhile can intelligibly doubt that the action is good (Moore 1903).

Originally, Moore attempts to use the open question argument to reject any kind of ethical naturalism. He seems to assume that concept distinction entails property distinction. If moral terms and natural terms have different meanings, then they cannot signify the same properties. But this premise is probably false. As contemporary philosophers generally believe, two different

concepts can still signify the same property. So, the open question argument is unable to argue against synthetic ethical naturalism. But I think that it can still reject analytic ethical naturalism such as analytic moral functionalism. Let's consider a candidate for the functional definition of 'rightness'. Suppose that 'being right' is, on this proposal, regarded as meaning 'causing an ideal observer to have the sentiment of approval'. The argument goes as follows:

(1) If two terms 'A' and 'B' have the same meaning, then the question "*x* is A, but is it B?" would be closed to a competent speaker.

(2) So, if the question "*x* is A, but is it B?" is an open question to a competent speaker, then 'A' and 'B' would have different meanings.

(3) The question "an action has the property of causing an ideal observer to approve of it, but is it right?" is an open question to a competent speaker.

(4) Therefore, 'being right' and 'causing an ideal observer to have the sentiment of approval' cannot have the same meaning.

Someone may point out that even a competent speaker might not recognize the truth of some complex analytic propositions, e.g. "triangular = trilateral". Since some analytic truths cannot be straightforwardly captured, they can be still intelligibly questioned. Thus, it could be argued that the open question argument is not a good test for analytic truths in general. The analytic moral functionalist can assert that although we can capture the functional roles of moral properties by conceptual analysis alone, we may not directly observe the identities of moral properties and certain functional properties, for functional expressions which define moral properties could be very complex, and hence be intelligibly questioned.

However, I want to add that *rational* competent speakers can still catch the truths of analytic sentences by conceptual analysis alone. By 'a rational competent speaker', I mean a person who grasps the meanings of concepts in question and, moreover, who has adequate logical and mathematical ability. Whereas a rational competent speaker can know a triangular object is a trilateral object solely by appeal to conceptual analysis and *a priori* reasoning, she cannot get the truth of "rightness is the property of causing an ideal observer to have the sentiment of approval" in this way. In other words, she might believe that an action can play the functional role R, but even after fully analyzing the meanings of the two concepts and even after doing logical and mathematical reasoning, she can still question intelligibly whether the action is right. Thus, we may revise the open question argument a little bit, but retain the original significance of the argument.

(1) If 'A' can be analytically defined as 'B', then the question '*x* is A, but is it B?' would be closed to a rational competent speaker.

(2) So, if the question "*x* is A, but is it B?" is an open question to a rational competent speaker, then 'A' cannot be analytically defined as 'B'.

(3) The question "an action has the property of causing an ideal observer to have the sentiment of approval, but is it right?" is an open question to a rational competent speaker.

(4) Therefore, 'being right' cannot be analytically defined as 'causing an ideal observer to have the sentiment of approval'.

The open question argument seems to apply to any other analytic naturalistic definitions of moral terms, so it is reasonable to assume that any version of analytic moral functionalism is also false. By contrast, synthetic moral functionalism denies that moral terms and functional terms have the same

meaning or purely conceptual connections, so it is immune to the open question challenge. According to synthetic moral functionalism, although moral terms and functional terms have different meanings, they still signify the same properties, just as 'water' and 'H_2O' denote the same object although they have different meanings—Moore is wrong in saying that concept difference entails property difference.

Atomistic vs. Holistic Moral Functionalism

Generally speaking, moral functionalism claims that moral properties are functional properties whose specification involves certain functional roles. In the preceding section, I have discussed the question of by what means we can capture the functional roles of moral properties. I have suggested that the functional roles of moral properties cannot be known without appealing to empirical investigation. In the current section, let's consider another question: What kind of functional roles moral properties involve? Are such functional roles *holistic* or *atomistic*? That is, can the functional roles of moral properties be characterized without reference to other moral properties? Most moral functionalists answer no.

To understand this point, we can have a look at the difference between functionalism and behaviorism in the philosophy of mind. Functionalism about the mind is generally regarded as a better alternative to behaviorism. Whereas behaviorism holds that mental states can be specified solely in terms of behaviors and behavior dispositions, functionalism asserts that we must appeal to other mental states in order to give an accurate account of what a mental state is. For

example, consider a behaviorist definition of the desire to eat ice cream. The behaviorist may say that a person has the desire to eat ice cream if and only if the person would go get a piece of ice cream if she sees it in a shop…. However, if the person believes that she has no money, she would not go get the ice cream, even though she has the desire. So, a more accurate account is: a person has the desire to eat ice cream if and only if the person would go get a piece of ice cream if she sees it in a shop *and* if she believes that she has enough money…. But this definition involves a mental state, i.e. this 'money belief'.[44] Now we can see why it is not adequate to define mental states without appealing to other mental states.

Most moral functionalists stress that the same is true of moral properties: in order to define what a moral property is, we must appeal to other moral properties. In other words, moral properties should be defined in a functional network involving other moral properties rather than be defined in an atomistic or linear manner—an example of atomistic functional definitions of moral properties is that "rightness is the property of promoting the overall pleasure".[45] Jackson and Pettit have a nice formulation of this idea:

> [W]e see no plausibility in the idea of providing an atomistic reduction of evaluative terms. Nor is the idea likely to appeal to many contemporary philosophers. By most accounts, moral terms are involved in a network of content-relevant connections, including connections with other moral and evaluative terms. No simple atomistic definition is going to yield an understanding of a moral term, because each such term is used in a way that presupposes a large network of connections with other terms, evaluative and descriptive (Jackson and Pettit 1995, p. 22).

---

[44] Similarly, in order to define what the money belief is, we have to involve other mental states.
[45] Some philosophers use 'functional properties' in the way that functional properties by definition have the holistic, networked character, whereas others use the term in a broad sense that accommodates non-holistic functional properties. Here I use the notion of 'functional properties' in the broad sense. See Shoemaker 1984.

Not only analytic moral functionalists adopt a networked definition of moral terms, but synthetic moral functionalists also notice the networked character of moral properties or normative properties. For example, Richard Boyd writes:

> In the particular case of the terms used to describe aspects of well-being this would entail that the natures to which they refer could be characterized not only in terms of features of the psychology of the individuals whose well-being is at issue, but also in terms of the relation between well-being and other morally relevant phenomena like goodness, fairness, sympathy, justice… in whatever combination is required by the task of explaining the achievements of moral practice (Boyd 2003, p. 28).

A holistic version of moral functionalism enjoys some advantages over a non-holistic or atomistic version. For example, holistic moral functionalism is in a better shape than its atomistic rival to preserve the normative significance of moral discourse. Atomistic moral functionalism as a kind of reductive ethical naturalism identifies, either analytically or synthetically, moral properties as purely natural properties without reference to other moral properties. For example, some version of atomistic moral functionalism might understand rightness as the property of maximizing the overall happiness. Philosophers usually hold that morality conceptually or intrinsically has a normative aspect. Suppose that you intend to do something right. Someone asks you: Why *ought* you to do this? You may answer: "Because this is right." The answer is perfectly legitimate. But if rightness is just the property of maximizing the overall happiness, your answer to that question would be: "Because this will maximize the overall happiness". The answer only indicates a descriptive fact. How can this descriptive fact that this action will maximize the overall happiness explain, on its own, why a person ought to perform the action? This seems to be a putative problem for atomistic moral functionalism (or any version of reductive ethical

naturalism). But if holistic moral functionalism is right that moral properties cannot be defined without reference to other moral properties, then the normativity of morality could be easily preserved. For example, when a holistic moral functionalist is asked the question "Why ought you to do the right action?", she can answer that it is because the action would be *fair*, or promote *good* things… I'm not claiming that reductive ethical naturalism definitely cannot solve the problem; I just maintain that this is a serious putative problem for reductive ethical naturalism, and the reductive ethical naturalist must take pains to solve this problem. Since reductive ethical naturalism *might* be defeated by this challenge, holistic moral functionalism involves less epistemic risk than atomistic moral functionalism.

Like most moral functionalists, I consider the networked character of moral properties as a fundamental feature of morality. In my dissertation, I will simply adopt the holistic version of moral functionalism without constructing more arguments to show this. But it seems that holistic functionalism is faced with a putative problem: if mental properties or moral properties can be only defined in terms of other properties in the same metaphysical category, there *appears* to be a circularity. In order to get out of this circularity, functionalists standardly adopt a logical technique called 'the Ramsey-Lewis method' to deal with networked definitions.

Originally, philosophers like Frank Ramsey and David Lewis developed the Ramsey-Lewis method to define theoretical terms in observational terms (Lewis 1970). Later some functionalists introduced the Ramsey-Lewis method into the philosophy of mind to functionalize mental properties. To see how this

method works, let me use the Ramsey-Lewis method to formulate a functional definition of, for example, 'pain'.[46]

First, develop a 'pain theory', T. For the sake of argument, let's suppose that T is an accurate and comprehensive theory that characterizes pain.

> (*T*) For any *x*, if *x suffers tissue damage* and **is normally alert**, *x* **is in pain**; if *x is awake*, *x* tends to be **normally alert**; if *x* **is in pain**, *x winces* and *groans* and **goes into a state of distress**; and if *x* **is not normally alert** or **is in distress**, *x tends to make more typing errors*. (The italicized expressions are nonmental predicates designating physical, biological and behavioral properties, states or events; the boldfaced expressions are psychological expressions designating mental properties.) (Kim 1996, p. 105)

Second, we can Ramsify *T* by existentially generalizing over each mental expression occurring in it, and hence we get a new theory, T_R:

> (T_R) There exists states M_1, M_2, and M_3 such that for any *x*, if *x suffers tissue damage* and is in M_1, *x* is in M_2; if *x is awake*, *x* tends to be in M_1; if *x* is in M_2, *x winces* and *groans* and goes into M_3; and if *x* is not in M_1 or is in M_3, *x tends to make more typing errors* (Kim 1996, p. 105).

Let's abbreviate T_R as '$\exists M_1, M_2, M_3 [T(M_1, M_2, M_3)]$'. It is important to note that T_R involves no mental constants. So, if we use T_R to define mental terms, there would be no circularity. Finally, we can get a functional definition of pain:

> pain = def. $\lambda x \exists M_1, M_2, M_3 [T(M_1, M_2, M_3)$ & *x* is in $M_2]$[47]

Similarly, we can use the Ramsification method to define moral properties. It is worth noticing that the Ramsification methodology is available to both analytic moral functionalism and synthetic moral functionalism. Before analytic moral functionalists (e.g. Jackson) introduced the Ramsey-Lewis method to define

---

[46] This example of Ramsey-Lewis method is from Kim 1996, pp. 105-6.

[47] '$(\lambda x)Fx$' has the same meaning as 'the property of being an *x* such that *x* is F'. Since mental properties are understood as second-order properties, pain is defined in the above way. A first-order functionalist like Lewis would define 'pain' as 'the $M_2 \exists M_1, M_2, M_3 [T(M_1, M_2, M_3)]$'. See Block 1980a.

moral terms, some synthetic moral functionalists had already adopted the method to account for moral properties (see Railton 1993). They differ only in whether the Ramsified definition is an analytic definition or a synthetic definition. Let's use the Ramsey-Lewis method to define moral terms as follows. First, consider an ethical theory of moral rightness, E:

> (*E*) For any *x*, if *x* is **good**, then *the distribution of x* that is **fair** is **right**; if *x causes the satisfaction of basic human needs*, *x* is **good**; if *x* is **right**, *x tends to cause flourishing of humans*. (The italicized expressions are descriptive predicates denoting descriptive properties, states, events, etc.; the boldfaced expressions are moral expressions designating moral properties.)

For the sake of argument, let's suppose that E is an accurate and comprehensive theory that characterizes rightness—of course, the real moral theory should be much more complex than E. Next, we can get another moral theory, E_R, by existentially generalizing over each moral expression:

> (E_R) There are three moral properties M_1, M_2 and M_3 such that for any *x*, if *x* is M_1, then *the distribution of x* that is M_3 is M_2; if *x causes the satisfaction of basic human needs*, *x* is M_1; if *x* is M_2, *x tends to cause flourishing of humans*.

Let's abbreviate E_R as '$\exists M_1, M_2, M_3 [E(M_1, M_2, M_3)]$'. Finally, we can get a functional definition of rightness without involving circularity:

> rightness = def. $\lambda x \exists M_1, M_2, M_3 [E(M_1, M_2, M_3) \& x \text{ is } M_2]$

Finally, I wish to indicate that the circle or network for functional characterizations should be wide enough. For instance, in scientific theories, scientific terms, such as 'mass', 'energy' and 'force', are inter-defined within a holistic network. This network is so wide as to involve many fundamental features, and hence advances our understanding of the world. But if the circle

were too small or tight, the definition by appeal to the circle would be uninformative and lack theoretical significance. For example, as the sensibility theorist suggests, to say that *x* is good, for example, is just to say that *x* is such as to make *appropriate* a particular sentiment of moral approval A (for normal humans as they now are and in normal circumstances), and to say that the sentiment A is appropriate is to say that it is a sentiment toward *good* objects in such and such conditions (McDowell 1997a, 1997b; Wiggins 1997). This circle is too tight and small to conduct interesting, significant information. But fortunately, a holistic version of moral functionalism can invoke a complex, wide network, which involves many significant phenomena, to define moral properties.

SSH Moral Functionalism and Non-reductive Ethical Naturalism

As I mentioned at the beginning of this chapter, I regard a holistic, second-order version of moral functionalism as a *non-reductive* form of ethical naturalism.[48] Richard Boyd also holds this view as I do. He writes:

> [T]he role associated with a (scientific, moral, theological…) term will involve inferential connections to lots of other terms associated with it in the same body of discourse. So, the defining nature associated with a term will ordinarily be specified in terms of a role which is *itself* specified by the relation between the term's referent and the referents of those other terms. There is thus ordinarily no prospect for turning the definitions provided by a naturalistic account of reference for terms within a domain into a *reductionist* treatment of the discourse within that domain… (original emphases) (Boyd 2003, p. 26)

[48] It is worth noticing that the holistic feature of moral functionalism alone doesn't entail the moral-descriptive irreducibility. If moral properties were functionalized as first-order realizers, moral properties would be reducible to descriptive properties, even though the functional networks of moral properties are holistic.

But I must admit that this claim is controversial and whether it is true depends on what we mean by 'non-reductive ethical naturalism'. Philosophers haven't discussed much the notions of non-reductive ethical naturalism. I will attempt to work on this in the following passages.

Some philosophers (e.g. Sturgeon and Jackson) seem to define non-reductive ethical naturalism in the below sense:

> (D1): Non-reductive ethical naturalism is the view that *a*) moral properties are themselves natural properties; *b*) moral properties are not identical with any properties which are signified by descriptive terms.

Let's turn back to holistic moral functionalism. If we use the Ramsification method to functionalize moral properties, we might signify moral properties by *descriptive* terms, i.e. the complex Ramsey expressions. According to D1, moral properties are reducible to descriptive properties anyway. Thus, on this view, holistic moral functionalism doesn't entail non-reductive ethical naturalism.

However, I very much doubt that this definition of 'non-reductive' is of much philosophical interest. After all, in order to know what a moral property is, we must rely on the relations between this moral property and other moral properties within a functional network characterized by the moral theory. In other words, the understandings of moral properties are always parasitic upon other moral or normative properties. Certainly, we can use the Ramsey-Lewis method to signify moral properties in descriptive terms, but how can a logical trick alone matter to the metaphysical issue of reducibility? In a word, it seems to me that Definition 1 fails to do justice to the well-grounded intuition that if holistic moral functionalism is true, then the moral domain would remain irreducible.

Philosophers of mind generally share the intuition with me. They believe that it doesn't follow from the mere fact that mental properties as functional properties can be signified by nonmental terms by the Ramsey-Lewis method that mental properties are reducible to nonmental properties or physical properties. A reductive physicalist must find other ways to argue for the reducibility of the mind. What reductive physicalists usually do is to argue that mental properties as functional properties are reducible to their first-order physical properties. Let me put this point in another way. It is generally agreed that if mental properties are functional properties, the multiple realizability of functional properties raises a *prima facie* challenge to reductive physicalism. No matter whether the challenge works eventually, reductive physicalists uniformly believe that they need to solve this problem. But *if* the Ramsey-Lewis method were to imply the reducibility of mental properties, it wouldn't matter whether mental properties are multiply realized by first-order physical properties, and then reductive physicalists would not worry about the 'multiply realizability' challenge. In the literature of the philosophy of mind, however, reductive physicalists have attempted to make *additional* arguments to show that multiple realizability is compatible with type physicalism.[49] Whether or not such arguments are tenable, the reason why reductive physicalists have tried to find other ways to establish the reducibility of the mind is precisely that they don't

[49] For example, some reductive physicalists argue that since there are no species-independent (or population-independent) mental states, multiply realizability doesn't threaten the identity of mental states and physical states-in-species. Or, some reductive physicalists suggest that even though mental properties are multiply realized by physical properties, mental properties are necessarily coextensive with the disjunctions of physical properties and are thereby reducible to physical properties.

believe that the Ramsey-Lewis method as a logical trick can solely establish mental-physical reducibility.

The same is true of moral properties. Even though we can signify moral properties by the Ramsey expressions that contain no normative terms, the 'real essences' of moral properties cannot be captured outside of moral domain, and hence moral properties still remain irreducible. For this reason, I suggest that we adopt another definition of non-reductive ethical naturalism:

> (D2) Non-reductive ethical naturalism is the view that *a*) moral properties are natural properties; *b*) moral properties are not identical with any properties which are investigated by descriptive natural disciplines.

D1 and D2 converge in most cases, since many properties that are signified by descriptive terms are also investigated by descriptive disciplines or theories. For example, "goodness is pleasure" is a case of reducibility according to both definitions. We can better know the difference between the two definitions when the two definitions go apart in some cases where a property signified by a descriptive term is not investigated by descriptive disciplines, such as the case of holistic moral functionalism. Although the Ramsey expression $\lambda x \exists M_1, M_2, M_3$ [$E(M_1, M_2, M_3)$ & x is M_2], for example, is a descriptive term, the understanding of the expression is parasitic upon moral or normative properties. Thus it is reasonable to say the property signified by the Ramsey expression is not a subject of any descriptive disciplines. So, the moral property rightness signified by the Ramsey expression is not yet reducible to any descriptive property on D2.

Someone might say that even if moral properties' being picked out by the (descriptive) Ramsey-Lewis expressions doesn't entail that moral properties are reducible to descriptive natural properties, we can find other ways to establish

the reducibility of moral properties. For example, some reductive ethical naturalists might argue that though multiply realizable moral properties are irreducible to any single first-order descriptive properties, they can be reducible to the disjunctions of first-order descriptive properties. This is actually a three-step argument. First, it is argued that moral properties are necessarily coextensive with the disjunctions of descriptive properties. Second, it is argued that necessary coextension entails identity. Third, it is argued that the disjunctions of descriptive properties are also *genuine* (descriptive) properties. Then the reductive ethical naturalist concludes that moral properties are identical with descriptive properties. I put the argument in a formal way below:

> (P1) Moral properties are necessarily coextensive with the disjunctions of descriptive properties.
>
> (P2) Necessary coextension entails identity.
>
> (P3) The disjunctions of descriptive properties are also genuine (non-normative) properties.
>
> (Conclusion) Therefore, moral properties are identical with descriptive properties.

In what follows, I will argue against P2 and P3, respectively. To begin with, I very much doubt that necessary coextension entails identity. For example, the property of being a round square is necessarily coextensive with the property of being a round triangle. But are they the same property? Probably not. According to the encoding account of property identity, the two properties are distinct properties, since they are not encoded by exactly the same abstract objects, even though they are necessarily coextensive with each other (see Zalta 1983).

Moreover, I want to introduce another counterargument against P2, which has a form of *reductio ad absurdum*. Let's leave aside for a moment the question of whether necessary coextension entails identity. It seems that necessary coextension is intuitively distinct from identity—it is reasonable to assume this.[50] Even philosophers who maintain that necessary coextension entails identity don't believe that the property of being necessarily coextensive is just the same relational property as the property of being identical. Rather, they just hold that necessary coextension is a sufficient condition for identity (that property *x* is identical with property *y* if *x* is necessarily coextensive with *y*). However, if necessary coextension were to entail identity (and, of course, identity entails necessary coextension), then the property of being identical and the property of being necessarily coextensive would be necessarily coextensive with each other. It follows that the two properties are the same property. But this contradicts our robust intuition that necessary coextension is distinct from identity. Thus, necessary coextension cannot entail identity.[51]

Now I want to argue against P3. I will attempt to show that the disjunctions of properties are themselves not genuine properties (unless they actually signify already-endorsed genuine properties). For example, the disjunctive property, being an even number or being an odd number, is actually the property of being a whole number, which is certainly a genuine property. So this disjunctive property, being an even number or being an odd number, should be also considered as a genuine property. Let's call disjunctive properties that

[50] Although it seems very clear to me that being necessary coextensive is a different (relational) property from being identical, I must concede that people who don't share the intuition with me won't regard this argument as a *reductio ad absurdum*.
[51] This point is inspired by Russ Shafer-Landau. See Shafer-Landau 2003, p. 91.

don't signify already-endorsed genuine properties 'merely disjunctive properties'. For instance, being white or being a table is probably a merely disjunctive property. In the current context, we don't need to worry about the possibility that the disjunctions of descriptive properties upon which moral properties supervene actually signify already-endorsed genuine properties. It has already been assumed that moral properties are multiply realized by descriptive properties. For instance, rightness is multiply realized by N_1, N_2, or N_3…. But if N_1 v N_2 v N_3… were actually another genuine descriptive property N*, then rightness would be *singly* realized by the descriptive property N*, which contradicts the assumption of multiple realization.[52]

Some philosophers seem to adopt an ultra-fine theory of property, according to which properties exactly correspond to concepts. Since there are disjunctive concepts, on this view, there are thereby (genuine) disjunctive properties. But most philosophers reject the ultra-fine theory of properties as introducing a superfluous ontology. Rather, it is generally held that for something to be a real property, it must be *projectible*. I will simply assume this theory of property in my dissertation without making any arguments.

But what is it for a kind to be projectible? Some philosophers, such as Pereboom and Kornblith, seem to maintain that something is a projectible kind *because* it can figure in laws:

[52] What I'm insisting is just that if moral properties are multiply realized by descriptive properties, those first-order descriptive properties cannot form a genuine *descriptive* kind; but I'm not sure if they cannot form any genuine kind. Suppose that rightness is a functional property and multiply realized by N_1, N_2, or N_3…. The disjunctive property, N_1 v N_2 v N_3…, may form a distinctive, genuine kind, 'a property that plays such and such functional role R'. However, if holistic moral functionalism is true, the functional role R cannot be specified descriptively; rather, role R can only be captured by reference to other moral or normative properties. So, this kind is actually an *irreducible* moral kind, according to my definition of non-reductive ethical naturalism (i.e. Definition 2).

> If kinds in psychology are multiply realizable… purported bridge principles relating psychological to neuro-physiological kinds will involve open-ended disjunctions…. Why are such disjunctions not natural kinds? Because they cannot appear in laws (Pereboom and Kornblith 1991, p. 126).

Other philosophers, however, seem to believe that projectibility is more primitive than lawlikeness—a kind cannot figure in laws *because* it is not projectible. For example, Fodor writes:

> [T]he biconditionals that connect functional properties with their realizers aren't *laws* (*a fortiori*, that they aren't 'bridge' laws). They *can't* be laws because the realizers of functional states are, by assumption, disjunctive and disjunctive properties are supposed to be *ipso facto* not projectible…. [D]efining the disjunction of two realizing states to be a single 'physical-chemical state' is not a metaphysical option that can be taken seriously (Fodor 1997, p. 155).

For our purposes here, I don't need to pursue the issue further. No matter which is more primitive than the other, it is reasonable to say that the fact that a putative property is unable to figure in laws provides an *epistemic* reason for showing that the putative property is not projectible (and hence that it is not a genuine property). So, I will attempt to show that merely disjunctive properties are not genuine properties by arguing that they cannot figure in laws.

There are various characteristic features of laws. For example, laws must support counterfactual conditionals. Also, laws must enable successful predictions. I wish to argue that although generalizations involving merely disjunctive properties may have *some* features of laws (such as counterfactual sustainability and predictability), they fail to exhibit a significant feature of laws: explanatory power. Laws must explain the occurrence of the events which instantiate their consequent; generalizations involving merely disjunctive properties are not genuine laws partly because they are not legitimate

explanatory theories (I have mentioned this in Chapter 4). Just as David Owens puts it,

> A law is confirmed by a pair of events which instantiate its antecedent and its consequent respectively if and only if the law explains the joint occurrence of these events.... [D]isjunctive laws neither explain nor are confirmed by their instances. I hope this goes some way towards accounting for the intuition that the disjunctive properties are 'unnatural' properties, properties which do not pick out any natural kind (Owens 1989, pp. 200, 202).

Pereboom and Kornblith bring forward a nice example to illustrate why a generalization involving merely disjunctive properties fails to play an (indispensable) explanatory role. Imagine that Mary walks down the street to buy an ice-cream cone. We can explain her behavior by appealing to the content of her beliefs and desires: she wanted an ice-cream cone and she believed one could be purchased down the street. But if we replace this explanation by one which contains an open-ended disjunction of physical predicates—if Mary is in state P_1 or P_2 or P_3, etc., she will move with trajectory T_1—then it seems that the new one is not a valid explanation (Pereboom and Kornblith 1991, p. 127). We must ask: What is common to and distinctive of P_1, P_2 and P_3... such that when Mary is in either state of them, she will behave so and so? To answer that these physical properties are all disjuncts of the disjunctive property is not interesting, on the one hand. To reply that all of them are realizers of certain mental states—so Mary will behave so and so when she is in either of them—fails to provide a purely physical explanation, on the other hand.

The same is true of moral explanations. If someone intends to explain why an 'unjust' country is in riot by saying that any country which is undemocratic or suffers from economic inequality or lacks political rights tends to be in riot (other

things being equal), such hypotheses appealing to disjunctive properties certainly have low-grade explanatory power and hence provide worse explanations than the hypothesis involving injustice. A multiply realizable property helps capture 'patterns' that the disjunctive property in question fails to capture—injustice provides some significant information that the disjunction of those descriptive realizers does not. In conclusion, since Premise 2 and Premise 3 are both problematic, the reductionist argument thus doesn't succeed. Then we can feel comfortable to adopt a non-reductive version of moral functionalism.

SSH Moral Functionalism Meets Moral Twin Earth (1)

In the previous sections, I have argued for a particular version of moral functionalism, i.e. SSH moral functionalism. First, SSH moral functionalism takes moral properties as second-order functional properties. Second, SSH moral functionalism is a kind of synthetic moral functionalism, according to which the functional natures of moral properties can only be discovered by appeal to empirical investigation. Third, SSH moral functionalism is holistic and hence antireductionist in the sense that the functional roles of moral properties cannot be specified without reference to other moral properties. In the rest of the chapter, I will consider two possible objections to SSH moral functionalism which focus on the last two characters, and attempt to make some responses to the two criticisms. The first objection aims at the causal semantics to which synthetic moral functionalism is usually committed, and the second objection attacks the holistic feature of my moral functionalism. The current section will discuss the first problem, and the next section will discuss the second one.

The new wave moral realists (including synthetic moral functionalists) standardly adopt a causal theory of reference. On this theory, the referents of proper names and natural kind terms are determined by a *causal* chain from the initial users to the current users; the referents are not determined by identifying descriptive content with the term in the speakers' head—although in order for a speaker to succeed in using a term to refer to what others earlier in the chain were using the term to refer to, the speaker must *intend* to do so. Speakers succeed in referring to something by means of the term because underlying their uses of the term are links in a causal chain stretching back to the dubbing of the object with that term. Some moral philosophers, such as Richard Boyd and Peter Railton, maintain that some version of causal semantics can also apply to moral terms (Boyd 1988; Railton 1989). Let me introduce Boyd's causal regulation theory. According to Boyd, a term *t* refers to a property P if and only if the property P uniquely causally regulates the use of *t*. What does he mean by 'causal regulation'? Boyd gives a detailed account as below:

> Roughly, and for nondegenerate cases, a term *t* refers to a kind (property, relation, etc.) *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k*.... Such mechanisms will typically include the existence of procedures which are approximately accurate for recognizing members or instances of *k* (at least for easy cases) and which relevantly govern the use of *t*, the social transmission of certain relevantly approximately true beliefs regarding *k*, formulated as claims about *t*..., a pattern of deference to experts on *k* with respect to the use of *t*, etc.... When regulations of this sort obtain, we may think of the properties of *k* as regulating the use of *t* (via such causal relations), and we may think of what is said using *t* as providing us with socially coordinated *epistemic access* to *k*; *t* refers to *k* (in nondegenerate cases) just in case the socially coordinated use of *t* provides significant epistemic access to *k*, and not other kinds (properties, etc.) (Boyd 1988, p. 116).

But *this* causal semantics of moral terms is faced with some serious challenges, especially the moral twin earth argument raised by Terry Horgan and Mark Timmons (see Horgan and Timmons 1991, 1992, 2000). Horgan and Timmons construct a thought experiment against synthetic ethical naturalism (including synthetic moral functionalism) equipped with the causal regulation theory. Horgan and Timmons ask us to imagine the moral twin earth scenario. Moral twin earth is just like earth in most respects. Moral twin earthlings behave like earthlings, and they speak a language which is very similar to English (let's call it 'Moral Twin English'). Moral twin earthlings also use 'good', 'right' to express their pro-attitudes toward certain actions, persons, characters, etc. and use 'bad', 'wrong' to express their negative attitudes toward certain objects (let's call such judgments which are very similar to earthling's moral judgments 'tworal judgments'). Moreover, moral twin earthlings also regard their tworal considerations as universalizable and overriding. But there is one crucial difference between earth and moral twin earth: whereas earthlings' use of moral terms is causally regulated in Boyd's sense (B-regulated, for short) by a certain set of natural properties, moral twin earthlings' use of tworal terms is B-regulated by a different set of natural properties. Then it follows from Boyd's causal regulation thesis that moral terms used by earthlings and tworal terms used by twin earthlings refer to different properties. So, they can't have genuine moral disagreement in beliefs; they just talk past each other. For example, suppose that the term 'right' in earth and moral twin earth is B-regulated by two different properties, E and T, respectively. So, according to Boyd's moral semantics, when earthlings claim that actions of lying (say) are right, they actually mean that lying

is E. And, when moral twin earthlings say "lying is wrong (or not right)", the sentence actually means that lying is not T. This apparent disagreement between earthlings and moral twin earthlings is just an illusion, from Boyd's point of view. But according to our robust intuition, earthlings and moral twin earthlings seem to engage in a real moral disagreement. Then, Horgan and Timmons insists, the new wave moral realism (also, synthetic moral functionalism) must be false.

However, it is important to note that Boyd's causal semantics is just a particular version of causal semantics and is somewhat different from other versions of causal semantics.[53] Furthermore, Boyd's causal semantics has some counterintuitive errors. If we can find a better version of causal semantics to solve the moral twin earth problem, the new wave moral realism could be still preserved. In order to see the defect of Boyd's causal semantics, let's consider an example raised by David Copp. Imagine some society in which earlier people intended to use the term 'honey' to refer to the kind of stuff that is actually honey. And suppose that there is a continuous, stable causal chain from the initial users to the current users. But despite this, people have astonishingly inaccurate and even bizarre beliefs about honey, such that their beliefs about honey are in fact hopelessly false and there is no tendency for their 'honey beliefs' to get any more accurate (Copp 2000). According to Boyd, whether a term *t* successfully refers to a kind *k* lies in whether the beliefs in virtue of *t* can be truer of *k* over time. So, in this case, 'honey' wouldn't refer to honey, on Boyd's view; perhaps it would refer to nothing. But this is quite counterintuitive!

[53] For example, Boyd's semantics is different from Putnam's semantics. According to Putnam, the referent of a term is determined by the first user's intention to pick out certain objects. But Boyd's semantics involves an epistemic element. On his view, whether a term *t* successfully refers to a kind *k* lies in whether the beliefs in virtue of *t* can be truer of *k* over time. See Copp 2000.

In what follows, I will introduce David Brink's counterfactual version of the causal regulation semantics. According to David Brink, the moral twin earth can be a serious problem for the new wave moral realism only when causal regulation is understood in a narrow, *actual* way; he instead proposes that we should understand causal regulation in a broad, counterfactual way. He writes:

> [The narrow interpretation] understands regulation in terms of the features of the world that causally regulate a speaker's actual use of moral terms. Determining which N-properties regulate a speaker's actual use of moral-kind terms looks like a historical inquiry rather than a moral inquiry, and the existence of significant moral disagreement suggests that actual usage is regulated by different N-properties in different speakers. However, causal regulation can and should be understood in counterfactual terms. On this view, terms refer to properties that regulate not just actual usage, but also counterfactual or hypothetical usage—in particular, the way speakers would apply terms upon due reflection in imagined situations and thought experiments (Brink 2001, p. 168).

According to this theory, a natural property N causally regulates a speaker's use of moral term 'M' just in case his use of 'M' would be dependent on his belief that something is N, were his beliefs in dialectical equilibrium.[54] Call a counterfactual causal regulation 'C-regulation'. Brink's version can correct the counterintuitive negatives implied by Boyd's semantics. According to Boyd, if our beliefs about honey are hopelessly false and there is no tendency for our 'honey beliefs' to get any more accurate, the term 'honey' doesn't refer to honey (it even refers to nothing). But if, as Brink suggests, the referent of 'honey' depends on our counterfactual beliefs about honey upon rational reflection—'honey' is C-regulated by honey, the term 'honey' can after all refer to honey.

---

[54] The counterfactual uses of moral term 'M' are actually dispositions of a particular kind characterized by counterfactual conditionals. The relata of causations can involve dispositions or instances of dispositions. So, the property N can be said to *cause* the disposition of using the term 'M' in a dialectic way; the counterfactual regulation in Brink's sense can also be regarded as a kind of causal regulation.

Now let's go back to the moral twin earth scenario. Although earthlings and moral twin earthlings actually hold two different theories, they can counterfactually have the same moral theory upon due reflection. In other words, although earthlings' and moral twin earthlings' moral terms are B-regulated by different sets of natural properties, they can be C-regulated by the same set of natural properties. Since the referents of moral terms are determined by the counterfactual moral theory, the actual difference in moral and tworal theories doesn't establish that earthlings and moral twin earthlings use moral and tworal terms to refer to different properties. Therefore, the moral twin earth argument fails to reject synthetic moral functionalism (or the new wave moral realism in general) by *only* attacking its moral semantics. But perhaps someone can use the moral twin earth argument or its variants to attack some particular features of SSH moral functionalism. I will discuss it in the next section.

SSH Moral Functionalism Meets Moral Twin Earth (2)

Although we can block Horgan and Timmons's argument against the moral semantics of SSH moral functionalism, SSH moral functionalism is still faced with another putative problem due to its holistic feature. I will discuss some counterargument against holistic moral functionalism and then attempt to show that those arguments cannot work. Since the alleged problem with holistic moral functionalism I will discuss later is actually a general (putative) problem with any other holistic versions of functionalism, such as functionalism about the mind and functionalism about meaning, I'm about to briefly discuss other versions of holism in order to see the putative problem more clearly.

Let's discuss functionalism about the mind first. As Ned Block indicates, if, according to functionalism, pain is defined in terms of certain functional roles within a causal network, this conception of pain would be too 'chauvinist'. Block puts it this way:

> Functionalists tend to specify inputs and outputs in the manner of behaviorists: outputs in terms of movements of arms and legs, sound emitted and the like; inputs in terms of light and sound falling on the eyes and ears. This conception is chauvinist, since it denies mentality to brains in vats and to paralytics. Such descriptions are blatantly species-specific. Humans have arms and legs, but snakes do not—and whether or not snakes have mentality, one can easily imagine snakelike creatures that do. Indeed, one can imagine creatures with all manner of input-output devices, e.g., creatures that communicate and manipulate by emitting strong magnetic fields (Block 1980b, p. 294).

Block's point is this. Suppose that the functional role of pain is characterized by a psychological theory P that specifies certain inputs and outputs. But it seems that we can conceive of a possible world in which the functional role of the creatures' pain state involves different inputs and outputs, which is characterized by a different psychological theory P*. Thus, functionalism about the mind, argues Block, is so chauvinist that it would (mistakenly) rule out the possibility that two instances of the same kind of mental state may involve different functional roles.

Then let me consider functionalism about meaning, i.e. functional role semantics. According to functional role semantics, the meaning or content of an intentional state (such as belief) is determined or type-individuated by the functional role of the intentional state, i.e. its *non-semantic* causal connections to (tokenings of) other mental states (Block 1986; Fodor 1987). But this semantic view seems to have some bizarre consequences, such as interpersonal incommunicability and intrapersonal instability. Since different people usually

don't have exactly the same cognitive system, it seems unlikely that two belief states in two persons' minds can have exactly the same functional role. So, on functional role semantics, probably no two belief states in different minds can have the same meaning. Hence people can hardly communicate with each other. Even worse, since belief states in the same mind at different times usually have different functional roles, it seems that any person cannot hold the same belief across time if functional role semantics is true. But this is obviously absurd! It seems to follow that the content of an intentional state is not determined by its functional role—functional role semantics is false.[55]

Now let's turn back to the moral domain. Likewise, if moral properties as functional properties are defined by their functional connections with other properties within the network, such conceptions of moral properties would also face the charge of chauvinism. Even though David Brink is right that the uses of moral terms should be causally regulated in a counterfactual way, it might be possible that people in different possible worlds would, even upon due reflection, come up with different normative theories that characterize different functional networks. Suppose that there are some differences between earth and moral twin earth. For example, the natural and social environments in the two planets are somewhat different. Or, people in the two planets might have slightly different cognitive and emotional systems. Suppose further that as a result the functional

[55] The functionalist may reply that the functional role of a belief, which is to determine the meaning of the belief, doesn't have to involve all connections between the belief and other mental states—rather, it only needs to involve some *subset* of the connections. Then, on this view, two belief states in different minds can have the same meaning so long as the two minds overlap to the extent that the two belief states have the same functional role specified by the same defining subset of connections. However, this reply just pushes the question back a step: Would two belief states *necessarily* have different meanings if they have different functional roles even understood in this sense?

roles of moral properties in the two planets are characterized by two different normative theories: in earth the functional network of moral properties is characterized by some specific consequentialist normative theory (call this theory T^c), while in moral twin earth the functional network of tworal properties is characterized by some specific deontological normative theory (call this theory T^d). But moral twin earthlings also use the term 'right', for example, basically as earthlings do: they intend to use the term to recommend certain actions, designate the property of 'to-be-doneness', and express certain positive attitudes etc. Also, moral twin earthlings regard their considerations about rightness as universalizable and overriding. According to moral functionalism (even equipped with a counterfactual version of causal semantics), the property denoted by the twin earth term 'right' is a different property from the one denoted by the earth term 'right'. Then earthlings and moral twin earthlings would talk past one another when they use 'right' to evaluate actions or policies. But it conflicts with our intuition that they do have a genuine moral disagreement. Therefore, moral functionalism *seems* to be false due to its holistic, networked feature, whatever moral semantics it adopts.

I have two responses. *First*, it is worth mentioning the general feature of holism that the values (the natures of moral properties, the meanings of belief, etc.) depend upon, or are a function of, certain functional networks. But we should notice that the general holistic claim need not commit one to the instability thesis. After all, one can claim that A is a function of B without implying that any change in B will produce a change in A. Let me introduce Henry Jackman's example. Consider, for instance, the claim that one's final letter

grade in a class is a function of, or depends upon, the results of one's exams, quizzes and homeworks. The truth of this claim certainly doesn't entail that no two people could have the same final grade unless they had precisely the same score on all of their homeworks, exams and quizzes. Neither does it entail that any change to one of one's quiz scores will produce a change in one's final grade. Each result makes *some* contribution to one's final grade, but not every change among the contributors will produce a corresponding change in the ultimate outcome. The function from contributing scores to final grades is *many-to-one* rather than *one-to-one*, and thus allows a good deal of stability in the value in spite of the possibility of tremendous variation in the variables (Jackman 1999). By the same token, although the natures of moral properties depend on their functional roles, i.e. the connections to other properties in a functional network, it doesn't logically follow that any change in the functional network will entail the change of the natures. In other words, it is possible that two different functional networks can define the same set of moral properties. Admittedly, I just showed that the moral twin earth argument doesn't *conclusively* reject holistic moral functionalism, since the same set of moral properties could be defined in terms of different functional networks. I haven't made a positive argument to show *how* the natures of moral properties *in particular* can remain the same within different functional networks—to be honest, I have no clue about how to argue for this in more detail.

So, let me introduce my second response. I intend to suggest that genuine moral properties across different possible worlds can be characterized by one and the same functional network (I can grant for the moment that different functional

networks entail different functional properties). As I mentioned above, Ned Block blames the functional definitions of mental states as 'blatantly species-specific'. But the functionalist can reply that it is just because the theory that is supposed to capture the functional roles of mental states is coarse or premature. A mature or ideal psycho-functional theory that aims to define mental states can be species-neutral or universal. The ideal theory would guarantee that the functional roles of mental states involve universal, species-neutral inputs and outputs to the effect that the functional definitions of mental states would accommodate, for example, Martian's pain, bat's pain, etc. I see no *a prior* reason why there cannot be such a theory that functionally defines mental properties across all possible worlds.

Similarly, the ideal moral theory that is used to functionally define moral properties should be adequately general and fundamental and involve universal descriptions of inputs and outputs (such as intentions, motivations, actions, etc.). This ideal moral theory should only capture the most fundamental and universal features of morality.[56] I found that Peter Railton makes a similar suggestion, when he talks about the functionalization of 'rationality':

> Among the clauses that state the job description should be some that have the effect of introducing slack, so that the definite description reads something like: "rationality is whatever unified (or almost unified) property plays all (or most) of these roles (or the most central among them) fully (or most fully)…". Without such a clause, the definition would badly over-commit, and not only about rationality…. We don't wish to represent our discourse or practices involving rationality as more definite than, in fact, they are (Railton 1993, p. 48).

[56] The ideal moral theory can be still synthetic or *a posteriori*.

So, holistic functional definitions *could* accommodate all genuine moral properties across all possible worlds, just as the 'H_2O' theory captures the essence of water in every possible world. I see no *a priori* reason why it cannot be so. And, if in some possible world (say, moral twin earth), the functional roles of putative moral properties are not captured by the ideal moral theory—for example, moral twin earthlings don't generally regard 'right' actions as being relevant to the promotion of well-being, or they don't usually feel guilty when they do something they call 'wrong'—then we have good reasons NOT to consider the properties denoted by such moral twin earth terms as genuine moral properties.

Actually Horgan and Timmons notice this kind of response. But they claim that it would lead to another problem, the problem of indeterminacy. Horgan and Timmons maintain that moral functionalism is impaled on two horns of a dilemma: the problem of conceptual relativism and that of indeterminacy. The first horn of this dilemma, which we have discussed, is that

> [I]f one grants that the proposed reference-fixing relation *R* suffices to pin down some unique class of natural properties as the putative referents of moral terms. We now tell a story about two groups, one on earth and one on Twin Earth where the natural properties *R*-linked to the moral terms as used by both groups are different on earth than on Twin Earth. Now the point of constructing this sort of scenario is that it reveals that the proposed version of naturalistic moral realism is committed to objectionable relativism.... The groups are really talking past one another; there is no genuine disagreement in the beliefs expressed by the assertions that employ the moral and the twin-moral vocabulary (Horgan and Timmons 2000, p. 140).

I have suggested that the functional roles of moral properties characterized by the ideal moral theory can apply to different groups in different possible worlds. So the challenge of relativism could be thus answered. But Horgan and Timmons

argues that this solution would be impaled on the other horn of this dilemma (i.e. the problem of indeterminacy). The second horn is that

> [T]he putatively reference-fixing relation *R* might fail to fix *determinate* reference-relations between moral terms and certain natural properties, because there are too many eligible natural properties that satisfy the constraints imposed by *R*. For instance, perhaps the *R*-constraints are satisfied by a class of natural properties—functional properties, say—that collectively satisfy some consequentialist moral theory T^c, and yet the *R*-constraints are *also* satisfied by another class of natural properties—also functional properties, say—that collectively satisfy some deontological moral theory T^d (Horgan and Timmons 2000, p. 140)

However, I can't help but ask: Do Horgan and Timmons here understand functional properties as second-order properties or first-order properties? Suppose that they regard moral properties as second-order properties and moral terms as second-order, rigid designators. If moral terms express second-order concepts and moral properties are second-order functional properties, then the referents of moral terms are completely determined by the ideal, universal moral theory T which characterizes the functional network in question. Since T is distinct from both T^c and T^d, how is it possible that functional properties captured by T^c and those captured by T^d can be eligible for being the referents of moral terms? Where does the charge of indeterminacy come from? Perhaps Horgan and Timmons understood functional moral properties as first-order realizers, and hence are arguing against first-order moral functionalism (there is also textual evidence to support this interpretation) (See Horgan and Timmons 2009). Horgan and Timmons's point may be this. If the ideal moral theory that is supposed to characterize the functional roles of moral properties is too universal or abstract, then different first-order natural properties would play the same functional roles. And if moral properties were considered as, on Horgan and

Timmons's view, first-order natural properties, then the ideal moral theory would fail to fix determinate referents of moral terms. Nevertheless, as I argued earlier, moral properties should be better understood as second-order properties. The natures of moral properties are thus completely determined by the functional network. Even though different first-order properties would realize the same functional roles in different possible worlds, moral properties as second-order properties would remain the same across different possible worlds. This is one advantage a second-order version of moral functionalism has over its first-order competitor.

In a nutshell, if the functional natures of moral properties are captured by the ideal, universal moral theory and if moral properties are second-order functional properties, moral functionalism would get out of the two horns of the putative dilemma.

Conclusion

As a version of moral functionalism, SSH moral functionalism enjoys the general advantages a functional account of moral properties has. For example, as I have argued in earlier chapters, moral functionalism can help account for the causal power of moral properties; moral functionalism can help vindicate the explanatory efficacy of moral properties; moral functionalism can explain, or cohere with, other widely-accepted theses (such as moral-descriptive supervenience and the multiple realizability of moral properties). Furthermore, as I have argued thus far in this chapter, SSH moral functionalism are more attractive than other versions of moral functionalism for several reasons. *First*,

because SSH moral functionalism is a version of synthetic rather than analytic moral functionalism, it is immune to the open question challenge. *Second*, since SSH moral functionalism understands moral properties as second-order functional properties, it can preserve the character of necessity of moral facts, and, moreover, help solve the moral twin earth challenge by dispelling the putative problem of indeterminacy. *Third*, SSH moral functionalism understands moral properties in a holistic, non-reductive way, so it can better preserve the normative significance of morality.

Bibliography

Antony, L. and J. Levine (1997) "Reduction with Autonomy", *Philosophical Perspectives* 11, pp. 83-105.

Armstrong, D. M. (1968) *A Materialist Theory of Mind*, London: Routledge & Kegan Paul.

Audi, R. (1997) "Ethical Naturalism and the Explanatory Power of Moral Concepts," in R. Audi, *Moral Knowledge and Ethical Character*, Oxford: Oxford University Press, pp. 112-28.

Audi R. (2003) *Epistemology: A Contemporary Introduction to the Theory of Knowledge*, New York: Routledge Press.

Blackburn, S. (1991) "Just Causes", *Philosophical Studies* 61, pp. 3-17.

Block, N. (1980a) "Introduction: What Is Functionalism?", in N. Block (ed.), R*eadings in the Philosophy of Psychology*, Vol. 1, Cambridge, MA: Harvard University Press, pp. 171-84.

Block, N. (1980b) "Troubles with Functionalism", in Ned Block (ed.), *Readings in the Philosophy of Psychology*, Vol. 1, Cambridge, MA: Harvard University Press, pp. 268-305.

Block, N. (1986) "Advertisement for a Semantics for Psychology", in P. A. French, T. E. Uehling, and H. K. Wettstein (eds.), *Midwest Studies in Philosophy: Studies in the Philosophy of Mind*, Vol. 10, Minneapolis: University of Minnesota Press, pp. 615-78.

Block, N. (1990) "Can the Mind Change the World?", in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge: Cambridge University Press, pp. 137-70.

Block, N. (2003) "Do Causal Powers Drain Away?", *Philosophy and Phenomenological Research* 67, pp. 110-27.

Bonjour, L. and E. Sosa (2003) *Epistemic Justification: Internalism vs Externalism, Foundations vs Virtues*, Oxford: Blackwell Publishing Press.

Boyd, R. (1988) "How to Be a Moral Realist?", in G. Sayre-McCord (ed.), *Essays in Moral Realism*, Ithaca: Cornell University Press, pp. 181-228.

Boyd, R. (2003) "Finite Beings, Finite Goods: The Semantics, Metaphysics and Ethics of Naturalist Consequentialism, Part II", *Philosophy and Phenomenological Research* 67, pp. 24-47.

Brink, D. (1984) "Moral Realism and Skeptical Arguments from Disagreement and Queerness", *Australasian Journal of Philosophy* 62, pp. 111-25.

Brink, D. (1989) *Moral Realism and the Foundations of Ethics*, Cambridge: Cambridge University Press.

Brink, D. (2001) "Realism, Naturalism, and Moral Semantics", *Social Philosophy and Policy* 18, pp. 154-76.

Burge, T. (1993) "Mind-Body Causation and Explanatory Practice", in Heil and Mele 1993, pp. 97-120.
Collins, J., Hall, N. and Paul, L. A. (eds.) (2004) *Causation and Counterfactuals*, Cambridge, MA: MIT Press.
Copp, D. (2000) "Milk, Honey, and the Good Life on Moral Twin Earth", *Synthese* 124, pp. 113-37.
Copp, D. (2003) "Why Naturalism?", *Ethical Theory and Moral Practice* 6, pp. 179-200.
Crisp, T. M. and T. A. Warfield (2001) "Kim's Master Argument", *Noûs* 35, pp. 304-16.
Cuneo, T. (2006) "Moral Facts as Configuring Causes", *Pacific Philosophical Quarterly* 87, pp. 141-62.
Darwall, S., A. Gibbard, and P. Railton (eds.) (1997) *Moral Discourse and Practice*, Oxford: Oxford University Press.
Dretske, F. (1993) "Mental Events as Structuring Causes of Behavior", in Heil and Mele 1993, pp. 121-36.
Eells, E. (1991) *Probabilistic Causality*, Cambridge: Cambridge University Press.
Fodor, J. (1965) "Explanations in Psychology", in M. Black (ed.), *Philosophy in America*, London: Routledge & Kegan Paul.
Fodor, J. (1974) "Special Sciences (Or: The Disunity of Science as a Working Hypothesis)", *Synthese* 28, pp. 97-115.
Fodor, J. (1987) *Psychosemantics*, Cambridge, MA: MIT Press.
Fodor, J. (1997) "Special Sciences: Still Autonomous after All These Years", *Philosophical Perspectives* 11, pp 149-63.
Friedman, M. (1974) "Explanation and Scientific Understanding", *Journal of Philosophy* 7, pp. 5-19.
Gibbard, A. (1990) *Wise Choices, Apt Feelings*, Cambridge, MA: Harvard University Press.
Gibbard, A. (2002) "Normative and Recognitional Concepts", *Philosophy and Phenomenological Research* 64, pp. 151-167.
Gibbard, A. (2003a) *Thinking How to Live*, Cambridge, MA: Harvard University Press.
Gibbard, A. (2003b) "Invoking Rationality to Explain Happenings", in J. L. Bermudez and A. Millar (eds.), *Reason and Nature: Essays in the Theory of Rationality*, Oxford: Oxford University Press, pp. 265-82.
Goldman, A. (1967) "A Causal Theory of Knowing", *Journal of Philosophy* 64, pp. 355-72.
Griesmaier, F. (2005) "Kitcher-style Unification and Explanatory Relevance", *dialectica* 59, pp. 37-50.
Harman, G. (1965) "Inference to the Best Explanation", *Philosophical Review* 74, pp. 88-95.
Harman, G. (1977) *The Nature of Morality: An Introduction to Ethics*, New York: Oxford University Press.
Harman, G. (1986) "Moral Explanations of Natural Facts—Can Moral Claims Be Tested Against Moral Reality?", *The Southern Journal of Philosophy* 24, Supplement, pp. 57-68.
Heil, J. and A. Mele (eds.) (1993) *Mental Causation*, Oxford: Clarendon Press.

Hempel, C. and P. Oppenheim (1948) "Studies in the Logic of Explanation", *Philosophy of Science* 15, pp. 135-75.
Hitchcock, C. (2002) "Probabilistic Causation", in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Stanford University.
Hitchcock, C. (2004) "Do All and Only Causes Raise the Probabilities of Effects?" in Collins, Hall and Paul 2004, pp. 403-18.
Horgan, T. (1989) "Mental Quausation", *Philosophical Perspectives* 3, pp. 47-76.
Horgan, T. and M. Timmons (1991) "New Wave Moral Realism Meets Moral Twin Earth", *Journal of Philosophical Research* 16, pp. 447-65.
Horgan, T. and M. Timmons (1992) "Troubles for New Wave Moral Semantics: The 'Open Question Argument' Revived", *Philosophical Papers* 21, pp. 153-75.
Horgan, T. and M. Timmons (2000) "Copping Out on Moral Twin Earth", *Synthese* 124, pp. 139-52.
Horgan, T. and M. Timmons (2009) "Analytical Moral Functionalism Meets Moral Twin Earth", in I. Ravenscroft (ed.), *Minds, Ethics, and Conditionals: Themes from the Philosophy of Frank Jackson*, Oxford: Oxford University Press, pp. 221-36.
Jackman, H. (1999) "Moderate Holism and the Instability Thesis", *American Philosophical Quarterly* 36, pp. 361-9.
Jackson, F. (1998) *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford: Clarendon Press.
Jackson, F. and P. Pettit (1990) "Program Explanation", *Analysis* 50, pp. 107-17.
Jackson, F. and P. Pettit (1992) "In Defence of Explanatory Ecumenism", *Economics and Philosophy* 8, pp. 1-21.
Jackson, F. and P. Pettit (1995) "Moral Functionalism and Moral Motivation", *Philosophical Quarterly* 178, pp. 20-40.
Kim, J. (1976) "Events as Property Exemplifications", in M. Brand and D. Walton (eds.), *Action Theory*, Dordrecht: Reidel Press, pp. 159-77.
Kim, J. (1984) "Concepts of Supervenience", *Philosophy and Phenomenological Research* 45, pp. 153-76.
Kim, J. (1993) "The Non-Reductivist's Trouble with Mental Causation", in Heil and Mele 1993, pp. 189-210.
Kim, J. (1994) "Explanatory Knowledge and Metaphysical Dependence", *Philosophical Issues* 5, pp. 51-69.
Kim, J. (1996) *Philosophy of Mind*, Boulder: Westview Press.
Kim, J. (1998) *Mind in a Physical World*, Cambridge, MA: MIT Press.
Kim, J. (1999) "Making Sense of Emergence", *Philosophical Studies* 95, pp. 3-36.
Kim, J. (2003) "The American Origins of Philosophical Naturalism", *Journal of Philosophical Research*, the APA Centennial Volume, pp. 83-98.
Kim, J. (2005) *Physicalism, or Something Near Enough*, Princeton: Princeton University Press.
Kim, S. (2000) "Supervenience and Probabilistic Causations", *Synthese* 122, pp. 245-59.
Kitcher, P. (1981) "Explanatory Unification", *Philosophy of Science* 48, pp. 507-31.
Kitcher, P. (1989) "Explanatory Unification and the Causal Structure of the

World", in P. Kitcher and W. Salmon (eds.), *Scientific Explanation*, Minneapolis: University of Minnesota Press, pp. 401-505.
Kornblith, H. (1994) "Naturalism: Both Metaphysical and Epistemological", *Midwest Studies in Philosophy* 19, pp. 39-52.
Leiter, B. (2001) "Moral Facts and Best Explanations", *Social Philosophy and Policy* 18, pp. 79-101.
Lewis, D. (1969) "Review of *Art, Mind and Religion*", *Journal of Philosophy* 66, pp. 22-7.
Lewis, D. (1970) "How to Define Theoretical Terms?", *Journal of Philosophy* 67, pp. 427-46.
Lewis, D. (1972) "Psychophysical and Theoretical Identification", *Australasian Journal of Philosophy* 50, pp. 249-58.
Lewis, D. (1973) "Causation", *Journal of Philosophy* 70, pp. 556-67.
Lewis, D. (1986) "Causal Explanation", in D. Lewis, *Philosophical Papers*, Oxford: Oxford University Press, pp. 214-40.
Lewis, D. (2000) "Causation as Influence", *Journal of Philosophy* 97, pp. 182-97.
Loeb, D. (2005) "Moral Explanations of Moral Beliefs", *Philosophy and Phenomenological Research* 70, pp. 193-208.
Majors, B. (2003) "Moral Explanation and the Special Sciences", *Philosophical Studies* 113, pp. 121-52.
McDowell, J. (1997a) "Values and Secondary Qualities", in Darwall, Gibbard and Railton 1997, pp. 201-14.
McDowell, J. (1997b) "Projection and Truth in Ethics", in Darwall, Gibbard and Railton 1997, pp. 215-26.
Moore, G. E. (1903) *Principia Ethica*, Cambridge: Cambridge University Press.
Nagel, T. (1978) "Ethics as an Autonomous Theoretical Subject", in G. S. Stent (ed.), *Morality as a Biological Phenomenon*, Berkeley: University of California Press, pp. 196-205.
Nozick, R. (1981) *Philosophical Explanations*, Cambridge: Belknap Press.
Owens, D. (1989) "Disjunctive Laws?", *Analysis* 49, pp. 197-202.
Pereboom, D. and H. Kornblith (1991) "The Metaphysics of Irreducibility", *Philosophical Studies* 63, pp. 125-45.
Putnam, H. (1973) "Reductionism and the Nature of Psychology", *Cognition* 2, pp. 131-46.
Putnam, H. (1975a) "The Mental Life of Some Machines", in H. Putnam, *Mind, Language, and Reality: Philosophical Papers*, Vol. 2, Cambridge: Cambridge University Press, pp. 408-28.
Putnam, H. (1975b) "The Nature of Mental States", in H. Putnam, *Mind, Language and Reality*: *Philosophical Papers*, Vol. 2, Cambridge: Cambridge University Press, pp. 429-40.
Railton, P. (1981) "Probability, Explanation, and Information", *Synthese* 48, pp. 233-56.
Railton, P. (1986) "Moral Realism", *Philosophical Review* 95, pp. 163-207.
Railton, P. (1989) "Naturalism and Prescriptivity", *Social Philosophy and Policy* 7, pp. 151-74.
Railton, P. (1993) "Noncognitivism about Rationality: Benefits, Costs, and an Alternative", *Philosophical Issues* 4, pp. 36-51.

Railton, P. (1995) "Subject-ive and Objective", *Ratio* 8, pp. 259-76.
Railton, P. (1998) "Moral Explanation and Moral Objectivity", *Philosophy and Phenomenological Research* 58, pp. 175-82.
Railton, P. (2005) "Moral Factualism", in J. Dreier (ed.), *Contemporary Debates in Moral Theory*, Oxford: Blackwell Publishing Press, pp. 201-19.
Rupert, R. (2006) "Functionalism, Mental Causation, and the Problem of Metaphysically Necessary Effects", *Noûs* 40, pp. 256-83.
Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.
Sayre-McCord, G. (1988) "Moral Theory and Explanatory Impotence", in G. Sayre-McCord (ed.), *Essays in Moral Realism*, Ithaca: Cornell University Press, pp. 256-81.
Sellars, W. (1968) *Science and Metaphysics*, London: Routledge & Kegan Paul.
Shafer-Landau, R. (2003) *Moral Realism: A Defense*, Oxford: Oxford University Press.
Shoemaker, S. (1984) "Some Varieties of Functionalism", in S. Shoemaker, *Identity, Cause, and Mind*, Cambridge: Cambridge University Press, pp. 261-86.
Shoemaker, S. (1998) "Causal and Metaphysical Necessity", *Pacific Philosophical Quarterly* 79, pp. 59-77.
Shoemaker, S. (2001) "Realization and Mental Causation", in C. Gillet and B. Lower (eds.), *Physicalism and Its Discontents*, Cambridge: Cambridge University Press, pp. 74-98.
Skyrms, B. (1980) *Causal Necessity*, New Haven: Yale University Press.
Sober, E. (1983) "Equilibrium Explanation", *Philosophical Studies* 43, pp. 201-10.
Sober, E. (1999) "The Multiple Realizability Argument against Reductionism", *Philosophy of Science* 66, pp. 542-64.
Street, S. (2006) "A Darwinian Dilemma for Realist Theories of Value", *Philosophical Studies* 127, pp. 109-66.
Sturgeon, N. L. (1985) "Moral Explanations", in D. Copp and D. Zimmerman (eds.), *Morality, Reason, and Truth: New Essays on the Foundations of Ethics*, Totowa: Rowman and Allanheld, pp. 49-78.
Sturgeon, N. L. (1986) "Harman on Moral Explanations of Natural Facts", *The Southern Journal of Philosophy* 24, Supplement, pp. 69-78.
Sturgeon, N. L. (1998) "Thomson Against Moral Explanations", *Philosophy and Phenomenological Research* 58, pp. 199-206.
Sturgeon, N. L. (2005) "Moral Explanations Defended", in J. Dreier (ed.), *Contemporary Debates in Moral Theory*, Oxford: Blackwell Publishing Press, pp. 241-62.
Sturgeon, N. L. (2006) "Ethical Naturalism", in D. Copp (ed.), *The Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press, pp. 91-121.
Suppes, P. (1970) *A Probabilistic Theory of Causality*, Amsterdam: North-Holland Publishing.
Thomson, J. (1996) "Moral Objectivity", in G. Harman and J. Thomson, *Moral Relativism and Moral Objectivity*, Oxford: Blackwell Publishing Press, pp. 65-154.

Thomson, J. (1998) “Reply to Critics”, *Philosophy and Phenomenological Research* 58, pp. 215-22.
Tye, M. (1983) “Functionalism and Type Physicalism”, *Philosophical Studies* 44, pp. 161-74.
Wiggins, D. (1997) “A Sensible Subjectivism”, in Darwall, Gibbard and Railton 1997, pp. 227-46.
Woodward, J. (2003) *Making Things Happen*, Oxford: Oxford University Press.
Yablo, S. (1992) “Mental Causation”, *Philosophical Review* 101, pp. 245-80.
Zalta, E. (1983) *Abstract Objects: An Introduction to Axiomatic Metaphysics*, Dordrecht: Reidel Press.
Zangwill, N. (2005) “Moral Epistemology and the Because Constraint”, in J. Dreier (ed.), *Contemporary Debates in Moral Theory*, Oxford: Blackwell Publishing Press, pp. 263-81.