RESPECT FOR PERSONS:
AN EPISTEMIC AND PRAGMATIC INVESTIGATION

by

Peter B. M. Vranas

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2001

Doctoral Committee:

Professor Stephen L. Darwall, Chair
Professor David O. Brink, University of California, San Diego
Professor Allan F. Gibbard
Associate Professor James M. Joyce
Professor Peter A. Railton
Professor Norbert W. Schwarz

ACKNOWLEDGMENTS

My greatest academic debt is to four members of my dissertation committee: David Brink, Allan Gibbard, Peter Railton, and especially Stephen Darwall. Each of them, over the last six years, gave me extensive comments on successive drafts of most parts of the dissertation. I am also deeply grateful to the other two members of my dissertation committee, James Joyce and Norbert Schwarz, with whom I discussed the dissertation less frequently but very profitably. My thanks go also to those current or former faculty members and students at the University of Michigan with whom I met occasionally to discuss parts of the dissertation: Elizabeth Anderson, Eugene Burnstein, Mark Crimmins, Edwin Curley, John Devlin, John Doris, Peter Gibbard, Rich Gonzalez, David Hills, Nadeem Hussain, Eric Lormand, Joshua Margolis, Andy Modigliani, Richard Nisbett, Gerhard Nuffer, Kevin Toh, David Velleman, and Kendall Walton. Thanks finally to many other people who asked interesting questions at talks, interviews, or discussions; I list them in footnotes at the beginning of individual chapters.

My greatest non-academic debt is to my parents: my father, who died while I was working on the dissertation, and my mother, whose unconditional love and support are still an important component of my life. I am also grateful to those of my friends from my former life as an engineer whose skeptical attitude towards philosophy helps me keep my activities in perspective.

PREFACE

We can distinguish (following Kant) two concepts of respect for persons: *appraisal respect* (or *esteem*), an attitude based on a positive appraisal of a person's moral character, and *recognition respect* (or *civility*), the practice of treating persons with consideration based on the belief that they deserve such treatment. After engaging (in Chapter 1) in an extended analysis of these concepts, I examine two "truisms" about them. (a) We justifiably believe of some persons that they have good (or bad) character and thus deserve our esteem (or contempt). (b) Frequently it pays to be disrespectful; e.g., insulting those who insult us may put them in their place. By using empirical results from social and personality psychology and techniques from decision theory in addition to conceptual considerations, I argue that, surprisingly, the above two "truisms" are false. (a) Extensive psychological evidence indicates that most persons are *indeterminate*—overall neither good nor bad nor intermediate—and that our information about specific persons almost never distinguishes those who are indeterminate from those who are not (Chapter 2). (b) The strategy of habitually avoiding disrespectful behavior maximizes long-term expected utility (Chapter 3). In sum, we have good pragmatic reason to treat persons respectfully, but we have good epistemic reason to avoid esteeming or despising them.

I wish to express the *hope* that no part of this dissertation will be quoted or referred to: although the dissertation as a whole is by no means rough, I intend to revise parts of it drastically before I submit them for publication in academic journals.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

RESPECT, *ACHTUNG*, AND SELF-RESPECT

'Respect' sometimes refers to *behavior* ("I did respect our agreement"), other times to a transitory *feeling* ("her handling of the crisis filled me with respect"), and still other times to a standing *attitude* ("my respect for him never wavered"). To avoid confusion, I will understand 'respect' simpliciter as "(having an) attitude of respect", which I will distinguish from "(having a) feeling of respect" and from "(engaging in) respect-related behavior".[1]

My central thesis in this chapter is that there are several distinct (though related) *kinds* of respect (and of contempt); i.e., attitudes to which 'respect' (or 'contempt') refers. My central task in this chapter is to examine these attitudes and their interrelationships. In §1 I elaborate on Darwall's (1977/1995) distinction between *appraisal* and *recognition* respect. In §2 I argue that a similar distinction is to be found in Kant's writings on *Achtung* ('respect'). In §3 I conclude with an examination of *self*-respect, including a reply to Dillon's (1992c) feminist critique of recognition self-respect.

---

[1] It seems that 'respect' unequivocally refers to behavior only when used as a *verb*. It also seems that, when used as a *noun*, 'respect' refers more frequently to an attitude than to a feeling (unlike, e.g., 'anger'). The three uses of 'respect' are related. For example, respect-related behavior is normally (at least partly) caused by an attitude or a feeling of respect. (Normally, but not invariably: respect can be faked.) Conversely, an attitude or a feeling or respect can (at least partly) cause respect-related behavior. (Can, but need not: respect can remain unexpressed.)

1. Appraisal versus recognition respect

1.1. Appraisal respect (AR) and appraisal contempt (AC)

Sometimes 'respect' has the sense of 'esteem', as in "I respect her for her integrity" (Frankena 1986: 155; Gibbard 1990b: 268; Hudson 1980: 72; Sachs 1981: 346). Esteem presupposes a positive appraisal; thus the label *appraisal respect*, coined by Darwall (1977/1995), seems appropriate for this kind of respect. Similarly, *appraisal contempt* presupposes a negative appraisal.[2]

The *object* of AR is the entity that is positively appraised. One can distinguish *personal* AR, whose object is a person, from *non-personal* AR, whose object is an entity other than a person (e.g., a performance, a character trait, an institution, or an animal).[3] One can also distinguish *moral* AR, whose object is appraised from a moral point of view, from *non-moral* AR, whose object is appraised from a non-moral point of view. Personal AR can be moral ("I respect you as a person") or non-moral ("I respect you as an athlete"); similarly, non-personal AR can be moral ("I respect the Administration's integrity") or non-moral ("I respect the Administration's efficiency"). I will understand 'AR' simpliciter as "personal moral AR".[4]

---

[2] The label *evaluative respect* (Hudson 1980; cf. Dillon 1992b) seems also appropriate. One could also use the label *estimative respect* (cf. Telfer 1968/1995); but "estimative contempt" sounds strange, whereas "appraisal contempt" sounds better.

[3] It seems possible to have some kind of esteem for institutions or animals (e.g., dogs or dolphins), contrary to Darwall's claim that the "exclusive objects" of AR are "persons or features which are held to manifest their excellence as persons or as engaged in some specific pursuit" (1977/1995: 184).

[4] Sometimes 'respect' has the sense of 'admiration' (Frankena 1986: 154). It's not admiration that I have in mind when I talk about non-moral AR. First, one kind of non-moral AR is clearly distinct from admiration: this is AR for persons "as engaged in some specific pursuit" (Darwall 1977/1995: 184). In order to be respected (rather than merely admired) as, e.g., a tennis player, it's not enough to be an excellent tennis player: one must also abide by standards of appropriate behavior that apply to tennis players qua tennis players (Darwall 1977/1995: 187). Second, even non-moral AR that is not AR for persons as engaged in some specific pursuit can be distinguished from admiration by noting that such non-moral AR cannot be grounded on merely natural abilities: "If someone is capable of some feat (which may be widely admired) solely by virtue of, say, his height, then neither this feat nor the person's ability to perform it are appropriate objects of respect" (Darwall 1977/1995: 188; cf. Cranor 1975: 312). Normally, however, it is "uncertain how great a part of the ability we admire must be ascribed to innate talent and how much to cul-

Following a widespread tradition in Social Psychology, I adopt a *tripartite* definition of attitudes, according to which an attitude in general consists of three components: cognitive, affective, and conative. I understand these three components as dispositions to have certain beliefs, affective reactions, and motives respectively. (See Appendix A for further details on this tripartite definition.) I will examine the three components of AR in turn.

1.1.1. Cognitive component

AR presupposes a positive appraisal of a person from a moral point of view; i.e., the person is appraised on the basis of her *character*.[5] One can distinguish *partial* AR, whose object is appraised on the basis of some aspect of her character, from *global* AR, whose object is appraised on the basis of an overall assessment of her character. Partial assessments of character can conflict: I can both respect you for your integrity (and thus have partial AR for you) and despise you for your stubbornness (and thus have partial AC for you). In cases of such a conflict I may still have global AR or global AC for you (if, e.g., I consider integrity much more important than stubbornness or vice versa); alternatively, I may have a *fragmented* attitude towards you, consisting neither in global AR nor in global AC.[6]

---

tivation through the person's own diligence" (Kant P5.078.05-06/T81/A138; see footnote 68 for an explanation of the way in which I refer to Kant's works), so that non-moral AR and admiration frequently go together and are thus sometimes confused. (Cranor uses "appraisal respect" in a *different* sense: he wants to leave open the possibility that "appraisal respect" might be grounded on "non-effort-related natural abilities or capacities, such as rationality or self-consciousness" (1982: 48; cf. 1983: 104).)

[5] According to Darwall (1977/1995: 188-90), the conception of character that is relevant to AR includes both (i) the higher-level disposition to do that which one takes to be supported by the best reasons and (ii) more specific dispositions (like honesty, fairness, kindness) to act for particular reasons. ("Higher-level" here is not used in Frankfurt's (1971/1986) sense, but refers rather to a *de dicto* disposition.) In Chapter II I argue that results from Social and Personality Psychology make appraisals of character epistemically unwarranted.

[6] Allan Gibbard gave me an example in which a fragmented overall attitude seems appropriate: when he was in Ghana, he heard about a doctor who had a clinic and was doing a lot of good in a backward and isolated part of the country, but had been a Nazi and had performed atrocities during World War II.

An appraisal is an assessment of value; in the case of AR, the value is moral goodness, goodness of character. A *positive* appraisal can be either an assessment of *positive* value (good character: *non-extremal* AR) or an assessment of *high* value (excellent character: *extremal* AR). The distinction between non-extremal and extremal AR, which is like the distinction between intelligence and brilliance, cuts across another distinction, namely the distinction between comparative and non-comparative AR. The object of *comparative* AR is appraised as having a better character than the "average" person (non-extremal comparative AR; cf. Rawls 1971: 437) or than the great majority of persons (extremal comparative AR) in a relevant population. The object of *non-comparative* AR, by contrast, is appraised as being a good person (non-extremal non-comparative AR) or an excellent person (extremal non-comparative AR) regardless of how many other persons are good or excellent.[7] It seems that AR vacillates between non-extremal and extremal, comparative and non-comparative AR. For example: is a merely "respectable" person (athlete, performance) respected? A positive answer suggests non-extremal AR; a negative answer suggests extremal AR.[8] In any case, all four kinds of AR admit of *degree*.[9]

---

[7] The distinction between comparative and non-comparative AR is different from the distinction between a comparative and a non-comparative conception of goodness. The latter distinction refers to whether goodness consists in having certain character traits (e.g., honesty) to a higher than average degree versus a given (independent of the average) degree. Having, e.g., a non-comparative conception of goodness is compatible both with comparative and with non-comparative AR. I want to leave open questions about how goodness and judgments of goodness are to be understood; thus my labeling beliefs about goodness "cognitive" is not intended to prejudge any issues about the proper analysis of moral judgments.

[8] Is it possible that everyone (or no one) be respected? If so, then we don't need the concept of comparative AR. A proponent of comparative AR can reply that, if everyone's character were to improve, then the standards that one would have to meet in order to be respected would correspondingly rise, so that not everyone would be respected. (People who were highly respected as marathon runners in 1900 might not have earned any respect as marathon runners by today's standards; cf. Nozick 1974: 241.)

[9] Comparative AR might also be called *other*-comparative. The object of *self*-comparative AR, on the other hand, is compared to oneself with respect to goodness of character, and is found to be better than oneself if self-comparative AR is *deferential*. It is this last kind of AR that Kant seems to have in mind when he talks about the "humble plain man, in whom I perceive a righteousness in a higher degree than I am conscious of in myself" (P5.077.01-03/T80/A136; cf. footnote 78). Note that AR which is neither other-comparative nor

1.1.2. Affective component

So far I have partially elucidated the cognitive component of AR. Now the affective component of AR consists in a disposition to have a positive affective response to the positive appraisal.[10] Such a disposition is arguably necessary for AR: a being lacking the ability to have the appropriate affective response[11] would probably not be said to have AR for anyone.[12] On the other hand, the disposition need not manifest itself for AR to exist: I need not lose my AR for you if I am temporarily depressed and thus don't have the appropriate affective response to my appraisal of you as a good person.[13] The last two examples suggest that one can *have* (the attitude of) AR without *feeling* AR, though probably not without a disposition to feel AR. Concerning the converse, I will argue that one can feel AR or have a disposition to feel AR without having AR.

I understand the feeling of AR not as "pure affect", but rather as having an affective, a cognitive, and possibly a conative component.[14] What, then, differentiates the

---

self-comparative still involves a comparison, namely between how a person *is* or behaves and how she *should* be or behave.

[10] It is important to note that the positive affect is a response *just* to the positive appraisal: if my breast swells with pride every time I think of my child's moral goodness, then I'm not responding *just* to the positive appraisal (but also, e.g., to the thought that it's *my* child), and I need not have AR. (I don't think it is necessary that the affective response be experienced as having a characteristic "feel" (e.g., by being recognized as similar to previous AR-related affective responses), but I remain neutral on whether such a condition is necessary in other cases (e.g., for fear).)

[11] By "lacking the ability" to have the affective response I understand something like, e.g., the case of persons who are congenitally unable to feel pain (Melzack & Wall 1988: 3-7).

[12] Stephen Darwall proposed the following putative counterexample to my claim that such a disposition is necessary for AR: if I believe on the basis of reliable reports that you are a good person, but every time I think of you I have a negative affective response to something bad you did to me during our single interaction many years ago, maybe I still have AR for you. But it seems to me that if I have AR, then I also have a disposition to have a positive affective response, even though this disposition never manifests itself because it is outweighed by a contrary disposition. So I don't think that we have a counterexample.

[13] Cf. Stocker (1979: 744) for cases of depression. Cf. also S. T. Coleridge, in *Dejection: An Ode*: "I see, not feel, how beautiful they are!" (Thanks to Peter Railton here.)

[14] Damasio (1994: chap. 7) emphasizes the cognitive component of feelings but also shows that the distinction between an affective and a cognitive component is to some extent artificial. Gibbard (1990b: 268

feeling of AR from (the attitude of) AR? There are at least three differences. First, attitudes normally last longer than feelings: as the example of depression suggested, the attitude of AR can persist even if the feeling of AR comes and goes. Second, the cognitive component of the feeling of AR is an *occurrent* belief (i.e., a *judgment*), whereas the cognitive component of the attitude of AR is a *dispositional* belief; i.e., a disposition to have an occurrent belief. (To see the difference, note that one may have a dispositional belief in the proposition that 28,367,072 is an even number without ever entertaining this proposition and thus without ever having a corresponding occurrent belief.) Third, even if a disposition to feel AR is necessary for having AR, it is not sufficient: I don't have AR if I take my feeling of AR to be unwarranted.[15] Such *subjectively unwarranted* feelings of AR can exist in at least two kinds of cases.

First: the cognitive component of the feeling of AR can be what may be called an *apparent* (as opposed to a real) occurrent belief. Even if it is my considered opinion (real belief) that you are a bad person, it can still *appear* to me as if you were a good person (e.g., because I'm in love with you). In this respect feeling AR is like feeling guilt: even if I believe that I have done nothing wrong and thus that I should not feel guilty, I can still feel guilty if it appears to me as if I had done something wrong.[16] Second: even if the cognitive component of my feeling of AR for you is a real belief, I can still believe that

---

n. 12) emphasizes the conative component of feelings: he wants to characterize feelings in part by their action tendencies.

[15] Strictly speaking, to have AR I must believe that *a* feeling of AR (not necessarily *my* feeling of AR) is warranted. For instance, if I believe that my feeling of AR is excessive and thus unwarranted, I can still have AR if I believe that a more moderate feeling of AR is warranted.

[16] Baumeister, Stillwell, and Heatherton (1994) make clear that the phenomenon of "guilt without transgression" is not uncommon. (1) "Receiving more than one deserves may cause guilt, especially in relation to other people who failed to be similarly overrewarded" (1994: 247). (2) "The phenomenon of survivor guilt ... makes clear that people can feel guilty without having done anything wrong, or indeed, having done much of anything at all. One simply feels guilty for surviving when others died" (1994: 251). (3) A study of unrequited love revealed that people who "became the reluctant agents of the would-be lover's heartbreak" felt guilty even when they had not encouraged the would-be lover and "regarded themselves as not having done anything wrong" (1994: 253; cf. Baumeister & Wotman 1992: chap. 5).

you don't deserve this feeling. For example, I can feel partial AR for you on the basis of my real belief that you are a good person insofar as you are exceptionally sincere, but take this feeling to be unwarranted because you frequently use your sincerity to hurt people's feelings. Some people may think that in this example I do have partial AR for you;[17] moreover, it does not seem possible to carry the example over to *global* AR. Be that as it may, the case of apparent beliefs is enough to establish that one can feel AR or have a disposition to feel AR without having AR.

Besides clarifying the relationship between the attitude and the feeling of AR, the above considerations elucidate further the cognitive component of AR in two ways. First: the positive appraisal consists in a real, rather than an apparent, dispositional belief.[18] Second: in addition to the positive appraisal, the cognitive component of AR includes the dispositional belief that a positive affective response to the positive appraisal is warranted,[19] in the sense of being deserved by the object of AR.[20] I will refer to the

---

[17] For instance, one might claim that I have both partial AR for you insofar as you are sincere and partial AC for you insofar as you hurt people's feelings. But this changes my example: I have in mind a case in which I have conflicting reactions to the *same* aspect of you, namely your sincerity. Or one might claim that there is some reason for feeling partial AR for you and some reason against feeling AR, so that I do have partial AR. (On my account having AR requires belief in the existence of *some* reason, not necessarily of *conclusive* reason, for feeling AR—see footnote 20). Again, this changes my example: I have in mind a case in which I feel partial AR for you while I believe that you *in no way* deserve this feeling. One might respond that in such a case I cannot believe that you are a good person insofar as you are sincere. I reply that only overall goodness, not partial goodness, may always deserve a positive affective response. (This reply, however, is not available in the modified example in which I have *global* AR: see next sentence in the text.)

[18] There is a subtle distinction between the case in which I have a subjectively unwarranted feeling based on an apparent belief (and thus I don't have AR) and the case in which I have a subjectively *warranted* feeling based on an apparent belief. The latter case corresponds not to AR, but to what may be called *appraisal acceptance*. For instance, I may be aware of your serious character flaws, but it may still appear to me as if your flaws didn't matter, and I may think that I should respond to this apparent belief as it if were real (e.g., because you are my spouse). Then I accept you despite (my awareness of) your flaws, although I don't have AR for you. (Note that, if I'm unaware of your flaws, then I don't accept you: I just have a mistaken real belief that you are a good person, so I can have AR for you.)

[19] The positive affective response is a response to the positive appraisal, and need not be caused by the belief that a positive affective response to the positive appraisal is warranted. Thus my account does not fall prey to "one thought too many"-style objections (Williams 1976).

latter belief as the cognitive$_2$ component of AR, and to the positive appraisal as the cognitive$_1$ component of AR.[21]

1.1.3. Conative component

Finally, the conative component of AR consists in a dispositional motive to have a positive behavioral response to the positive appraisal.[22] (A dispositional motive is a disposition to have an *occurrent* motive;[23] the latter is the conative component of the *feeling* of AR.[24]) The behavioral response can take a number of forms (Cranor 1975: 314-5; Maxwell & Silverman 1978): one may emulate the object of respect; in the case of extremal (moral or nonmoral) AR, one may honor her (e.g., by bowing, saluting, using honorific titles, or awarding prizes); in the case of nonmoral AR, one may defer to her opinion (if one respects her as an expert); and so on.

---

[20] The clarification that warrant is understood as based on desert is essential: I don't have AR if I just have some self-interested reason for feeling AR. Note also that, to have AR, one need *not* believe that a positive affective response to the positive appraisal is warranted in the sense that one has *conclusive* reason to have such a response. If you are my best friend's wife and I feel, to my dismay, that I'm falling in love with you, then I may have conclusive reason to suppress my feeling of AR for you, even if I believe that you deserve this feeling.

[21] A further clarification about the cognitive$_1$ component of AR: the positive appraisal must consist in the belief that the object of AR is a good person, not in the belief that she has certain character traits (which in fact make her a good person). An amoralist, for instance, who has a positive affective response to a sincere person just because she happens to like sincere people (rather than because she believes that sincerity is good) does not have AR for this person. This requirement guarantees (but is not guaranteed by) Darwall's requirement that AR "must be a *categorical* attitude, one which is unconditional on the fact that the traits in question happen to serve some particular purpose or interest of *mine*" (1977/1995: 190). (I don't intend to exclude overdetermination, and I guess neither does Darwall: if the fact that your traits serve some interest of mine contributes to my positive affective response but is not essential for it, then I may still have AR.)

[22] Strictly speaking, it's not the positive behavior, but rather the disposition (to have an occurrent motive to engage in positive behavior) that is a response to (i.e., is caused by) the positive appraisal. (More on what is "positive" behavior in §1.2.1.)

[23] Following the "Humean theory of motivation" (M. Smith 1994: chap. 4), I understand an occurrent motive as consisting (roughly) in a desire and a means-ends belief.

[24] Note that, in the case of the feeling, the response may be to an *apparent* belief.

Component	Attitude of AR	Feeling of AR
Cognitive$_1$	Positive appraisal; i.e., real dispositional belief that the object has (positive or high) moral value (goodness of character).	Real *or apparent* occurrent belief that the object has moral value.
Cognitive$_2$	Dispositional belief that a positive affective response to the positive appraisal is warranted (i.e., deserved by the object).	___
Cognitive$_3$	Dispositional belief that a positive behavioral response is warranted.	___
Affective	Disposition to have a positive affective response to the positive appraisal.	Positive affective response to the (real or apparent) occurrent belief that the object has moral value.
Conative	Dispositional motive to have a positive behavioral response to the positive appraisal.	(Possibly:) Occurrent motive to have a positive behavioral response to the (real or apparent) occurrent belief that the object has moral value.

Table 1.1. Cognitive, affective, and conative components of the attitude and of the feeling of AR

In analogy with what I said when discussing the affective component of AR, note the following points. A disposition to have some appropriate motive is arguably neces-

sary for AR.[25] Such a disposition, however, need not manifest itself (in an occurrent motive) for AR to exist. (Here we have a difference between the affective and the conative component of the *feeling* of AR: one cannot feel AR without positive affect, but one can feel AR without any occurrent motive.) Finally, the cognitive component of AR includes the dispositional belief (cognitive$_3$ component) that the object of AR deserves a positive behavioral response.

Table 1.1 summarizes the main conclusions of the discussion so far. Briefly (but roughly), having AR amounts to having both the disposition to feel AR and the dispositional belief that this disposition is warranted.[26]

1.1.4. Mistakes

The concept of AR can be further clarified by thinking about *mistakes* related to AR. I will briefly examine four kinds of mistakes: causal, conceptual, factual, and normative. (1) If I believe that my positive affect towards you is caused by my positive appraisal of you, whereas the affect is caused, e.g., by the fact that you unconsciously remind me of my mother, then I'm making a *causal* mistake. In such a case I don't feel AR, although I think I do. (2) If I feel love but not AR for you but I call my feeling "esteem" because I'm confused about the difference between love and esteem (as anecdotal evidence suggests people sometimes are), then I'm making a *conceptual* mistake and, again, I don't feel AR, although I think I do. (3) If I feel AR for you because I believe that you are honest whereas in fact I'm your dupe, then I'm making a *factual* mis-

---

[25] Thus I disagree with Darwall's assertion that AR for a person "does not essentially involve any conception of how one's behavior is appropriately restricted" (1977/1995: 186).

[26] I say "roughly" partly because a disposition to have a positive affective response to an occurrent belief differs from a disposition to have a positive affective response to a dispositional belief. It may be argued (contrary to Table 1.1) that the affective component of AR is the former disposition, because an affective response cannot be caused by a disposition (to have an occurrent belief). I want to leave this issue open, because what causes what here seems to be an empirical rather than a conceptual matter.

take. In such a case my feeling of AR is objectively unwarranted; if, however, my feeling is subjectively warranted, then I do have AR. (4) If you are a bad person but I appraise you positively because I have a wrong idea of what a good person is, then I'm making a *normative* mistake. A pirate, for example, may be aware of her captain's cruelty towards civilians but believe that the captain is a good person because the captain is highly considerate towards her crew. In such a case, again, the pirate does feel (and have) AR, although her feeling is objectively unwarranted.[27] Cases like (3) and (4) suggest that the components listed in Table 1.1 are jointly sufficient for feeling or having AR; cases like (1) and (2) suggest that these components are severally necessary. In conclusion, it seems that I have provided necessary and sufficient conditions for having and for feeling AR.[28]

Concerning finally appraisal *contempt*, I have very little to say. If one substitutes 'negative' for 'positive', then the discussion summarized in Table 1.1 seems to apply, mutatis mutandis, to AC as well.

1.2. Recognition respect (RR) and indifference contempt (IC)

The (plausible) claim that everyone is entitled to respect is different from the (paradoxical) claim that everyone is entitled to esteem; thus 'respect' does not always have the sense of 'esteem', and AR is not the only kind of respect (Frankena 1986: 155; Sachs 1981: 346). The central component of AR was the positive appraisal; by contrast, the central component of what I call (following Darwall 1977/1995) *recognition re-*

---

[27] This case should be distinguished from a case in which one judges a person to be bad but one has a positive affective response to this judgment. If this new case is at all intelligible, it involves no (feeling of) AR, because it involves no (real or apparent) belief that the person is good.

[28] My claim that I have provided necessary and sufficient conditions for feeling AR should be distinguished from the claims that there are certain symptoms which always accompany the feeling of AR or that there are certain situations which always cause AR; I'm not making the latter claims.

spect[29] is the dispositional belief that a certain behavior towards the object of respect is *deserved* by the object, in the sense of being warranted solely by virtue of features of the object.[30] RR for the government, for instance, usually presupposes the dispositional belief that the government deserves to be obeyed, maybe because it was democratically elected. RR for the government can coexist with a negative appraisal of the government as being corrupt, and thus with AC for the government.

Like AR, RR can be *personal* or *non-personal* (according to whether the object is a person or, e.g., an institution, a decision, a treaty, or the environment),[31] and *moral* or *non-moral* (according to whether the object's desert is taken to provide one with a moral or with a non-moral—e.g., a prudential—reason for behaving in a certain way[32]).[33] For the moment I will *not* restrict my attention to personal moral RR. One can also distinguish *non-specific* from *specific* RR, according to whether the object's desert is taken to provide one with a reason for behaving in *some* (unspecified) way versus a (more or less) specific way. For example, a boxer who respects an opponent's left hook (Darwall 1977/1995: 186) in the sense of (essentially) believing that caution is called for on the

---

[29] Cranor (1982: 53-5, 1983: 108-9) and Frankena (1986: 156-7) use 'consideration respect' synonymously with 'recognition respect'.

[30] "Desert can be ascribed to something or someone only on the basis of characteristics possessed or things done by that thing or person. That is, desert is simply never forward-looking" (Kleinig 1971: 72). Cf. Falls (1987: 39): desert is "non-consequentialist" (cf. also Cranor 1975: 311). (See footnote 41 for the contrast that I make between desert and merit.)

[31] Non-personal RR is wider than non-personal AR, in the sense that some possible objects of non-personal RR are not possible objects of non-personal AR (but not vice versa): e.g., one can have RR (but not AR) for a dangerous substance, in the sense of (essentially) believing that caution is called for when handling the substance. (RR for a dangerous substance is a case of *obstacle respect* (Hudson 1980: 74; cf. Feinberg 1973: 1).)

[32] Note that this definition of the moral/non-moral distinction for RR differs from the corresponding definition for AR (§1.1).

[33] As Kleinig points out, "desert is not a specifically moral notion. Although desert claims often—perhaps usually—have moral overtones there does not seem to be any necessity that they should" (1971: 72). For instance: "We can quite properly speak of the Niagara Falls being deservedly famous or of the Western Australian coastline deserving to be as well known as that of the East" (1971: 71-2).

ring, may have no idea how to handle the opponent's left hook (non-specific RR), or may have some more or less detailed strategy in mind (specific RR). Specificity admits of degree, and non-specific RR might be just an extreme case of specific RR (at the low end of the specificity "scale").[34] Even so, the concept of non-specific RR is of interest in its own right, as shown by a contrast with what I call *indifference contempt*.

The central component of IC is the dispositional belief that there is *no* way in which the object of IC deserves to be treated. This belief should be distinguished from the belief that there is no way in which the object of IC *ought* to be treated:[35] if I believe that I ought to water your plant just because I promised you to do so (rather than, e.g., because I believe that the plant deserves to be preserved), then I can still have IC for your plant. Having (personal moral) IC for someone presupposes the belief that, as far as her deserts go, it is morally indifferent how she is treated. Thus IC is in a sense an *extreme* form of contempt: it amounts, essentially, to viewing people like, e.g., rocks (Frankena 1986: 153; Norman 1989: 329). The contrast between the concepts of IC and of non-specific (personal moral) RR shows that the latter concept is useful insofar as it captures an attitude of *minimal* respect for persons.

---

[34] The contrast between specific and non-specific RR is like the contrast between what Quine (1955/1976: 185-6) calls the *relational* and the *notional* sense of (e.g.) sloop-wanting. Quine's contrast, however, becomes less clear-cut once it is realized that specificity admits of degree: as the set of specific sloops among which my preference lies gets bigger and bigger, my desire for a sloop arguably approaches (but does it ever reach?) a desire for "mere relief from slooplessness".

[35] (Regardless of whether "ought" is moral or prudential, *pro tanto* or all-things-considered.) To see more clearly the contrasts between specific and non-specific RR and between non-specific RR and IC, consider the following rough formalizations. Let o be an object and t range over treatments. (1) If I have specific RR for o, then $\exists t$ (I believe (o deserves t)). (2) If I have non-specific RR for o, then I believe ($\exists t$ (o deserves t)). (3) If I have IC for o, then I believe ($\neg\exists t$ (o deserves t)). The fourth case, $\neg\exists t$ (I believe (o deserves t)), seems to correspond neither to respect nor to contempt.

1.2.1. Positive respect (PR) and negative contempt (NC)

Even though the concept of RR is useful insofar as RR excludes one kind of contempt, namely IC, the concept of RR is too wide insofar as RR *includes* another kind of contempt. If I believe that infidels deserve to be tortured or exterminated, then I can have RR for infidels but it sounds strange to say that my attitude is a kind of respect. It seems that respect normally presupposes a belief that some *positive* treatment is warranted, whereas desert is compatible both with positive and with negative treatment: one can deserve a reward or a punishment. Thus one can distinguish two kinds of RR, which I call *positive respect* (PR) and *negative contempt* (NC), according to whether one believes that the object of RR deserves positive or negative treatment respectively. NC has the paradoxical feature of hovering between respect and contempt (this is why I call it negative *contempt*, although it's a kind of recognition *respect*): in some cases NC is more like contempt than like respect (e.g., NC for infidels),[36] but in other cases NC is more like respect than like contempt (e.g., if I believe that you deserve to be punished for some wrongdoing; cf. Kant M6.333.21-22/T106). The fact that recognition respect, as usually defined (Cranor 1982: 53-4, 1983: 108; Darwall 1977/1995: 183, 185; Dillon 1992a: 111, 1992d: 72; Frankena 1986: 156-7), includes NC seems to have escaped notice in the literature. I suspect that some of the authors who talk about "recognition respect" may have something like PR in mind—or may be vacillating between RR and PR.

What constitutes *positive* treatment? There are two potentially competing answers (cf. Atwell 1982: 22-5; Noggle 1999: 449-50). On one interpretation, treating an object positively amounts to promoting the object's good. This is possible for *some* kinds of objects (e.g., persons or plants) but not for others (e.g., contracts or materials). On another interpretation, treating an object positively amounts to abiding by the object (e.g., a

---

[36] If one believes that torturing infidels saves their souls and is thus positive treatment, then one has PR (rather than NC) for infidels, although (I take it) one is making a normative mistake (cf. §1.1.4).

treaty), to avoiding coercion or interference (e.g., respecting someone's decision).[37] A conflict arises when promoting an object's good is taken to require coercion or interference, as in the case of standard paternalistic laws (e.g., laws mandating seat-belt use or prohibiting drug consumption; cf. Dillon 1992a: 125-6). It is not my purpose here to resolve such conflicts: my task in this chapter is descriptive rather than normative, so I am simply drawing attention to an ambiguity concerning "positive" treatment, and thus to a tension between two views on the behavioral content of PR (and of NC). Attending to this tension helps to explain how "recent practical moralists have made appeal to respect for persons for the sake of supporting almost every imaginable policy—capital punishment *and* its abolition, abortion on demand *and* the 'right to life'," and so on (Atwell 1982: 17; cf. Edel 1974). Some philosophers contrast respect with love and focus on noninterference when explicating respect. According to Kant, in observing a duty of respect "I keep myself within my own bounds", and the principle of respect admonishes people "to keep themselves *at a distance* from one another".[38] Similarly, Norman claims that respect "involves a sense of *separateness* from others", "is a reaction of distancing oneself, recognising that the other person's projects are his, not mine. It is an inclination, not to live the other's life for him, but to stand aside and let him live his own life in his own way" (1989: 325-6). (Cf.: Ghosh-Dastidar 1987: 84-5; Gibbard 1990b: 237-8, 264-9.) Other philosophers, arguably a minority, seem to equate treating positively with promoting the object's good. Maclagan claims that in order to respect "we have to make

---

[37] The distinction between positive and negative treatment seems inapplicable to some kinds of objects (e.g., an opponent's left hook or a dangerous substance). Thus PR and NC don't exhaust RR; the remaining part of RR seems to correspond to obstacle respect (footnote 31). (Allan Gibbard suggested to me that 'respect' is used metaphorically in expressions like "respecting an opponent's left hook" or "respecting the power of the storm" (cf. Sachs 1982: 114). Feinberg (1973: 1), on the other hand, claims that the sense of 'respect' in examples like the above is etymologically prior to other senses of 'respect'.)

[38] M6.449.10-11,6.450.10-11/T198,199. Cf.: "love can be regarded as attraction and respect as repulsion, and if the principle of love bids friends to draw closer, the principle of respect requires them to stay at a proper distance from each other" (Kant M6.470.04-07/T215).

others' ends our own" (1960b: 294), and comes close to saying that "Agape" and respect for persons are the same thing (1960a: 216; cf. Harris 1966: 113-4). Similarly, Dillon (1992a, 1992d) argues that one kind of respect (which she calls "care respect") comes close to a kind of love. (Cf. Donagan 1977: 65.)[39]

1.2.2. Legitimacy respect (LR) and merit respect (MR)

So far I have elucidated what (in analogy with my terminology for AR) can be called the cognitive$_3$ component of RR, namely the dispositional belief that the object deserves a certain treatment. One cannot believe that a treatment is deserved without also believing that there is some reason why the treatment is deserved (Feinberg 1963/1970b: 58; Kleinig 1971: 73; Pojman & McLeod 1999: 62);[40] I will call the dispositional belief that this reason obtains the cognitive$_1$ component of RR. One can distinguish two kinds of RR, according to whether this reason is taken to be that the object has legitimacy (*legitimacy respect*, LR) or merit[41] (*merit respect*, MR). To see the difference between LR and MR, take an example. A suggestion about whether a faculty member should be promoted may deserve consideration for either of two reasons: the suggestion may have legitimacy (e.g., because it's a recommendation by a specially appointed panel of experts), or the suggestion may have merit (e.g., because it's backed up by good argu-

---

[39] I'm *not* claiming that philosophers who focus on noninterference when explicating respect think that interference is never justified. Some such philosophers even think that interference is sometimes justified *by* considerations of respect, rather than just *compatibly* with such considerations (Brody 1982; Gruzalski 1982). (So I'm not claiming that every philosopher who thinks that respect sometimes involves interference also thinks that treating positively amounts to promoting the object's good.)

[40] The latter belief can be understood non-specifically (one may be uncertain about what the reason is) or specifically (one may have a particular reason in mind); in what follows I consider only the specific understanding.

[41] Although 'merit' and 'desert' are sometimes used interchangeably, I understand merit as a *value* that *grounds* desert (cf. Sher 1987: chap. 7). Unlike (e.g.) Dillon (1997: 229), I don't understand merit as goodness of character.

ments).[42] Disagreement may exist about what grounds legitimacy or merit: is a recommendation by a committee of undergraduate students legitimate or should undergraduates have no say in the promotion process? One can have LR without having MR and vice versa: one can believe that the experts' recommendation has legitimacy but is based on sloppy reasoning and thus has no merit, and one can believe that the undergraduates' recommendation lacks legitimacy but is based on careful reasoning and thus has merit.[43]

<table>
<tr><th colspan="3">Recognition respect (RR)</th><th colspan="2" rowspan="2">Indifference contempt (IC)</th></tr>
<tr><th>Positive respect (PR)</th><th colspan="2">Negative contempt (NC)</th></tr>
<tr><td>Positive legitimacy respect, Positive merit respect</td><td>Negative legitimacy respect, Negative merit respect</td><td rowspan="2">Illegitimacy contempt, Demerit contempt</td><td>Defiance contempt (legitimacy, merit)</td><td>Disregard contempt (nonlegitimacy, nonmerit)</td></tr>
<tr><td colspan="2">Legitimacy respect (LR), Merit respect (MR)</td><td colspan="2"></td></tr>
</table>

Table 1.2. Kinds of recognition respect and of indifference contempt

Table 1.2 summarizes the three main distinctions drawn so far, namely: (1) the distinction between RR and IC (is there or not some treatment that the object is taken to deserve?); (2) the distinction between PR and NC (is the object of RR taken to deserve positive or negative treatment?); and (3) the distinction between LR and MR (is the ob-

---

[42] The belief that the suggestion is legitimate does not amount to the belief that the suggestion deserves consideration: the suggestion deserves consideration *because* it is legitimate. But does the former belief have the latter belief as a consequence? No, because disagreement may exist about the response that legitimacy makes appropriate: should *every* legitimate suggestion be discussed in a faculty meeting? But can a belief about legitimacy exist without *some* specific belief about appropriate treatment? Yes: one can be uncertain about what treatment is appropriate. Maybe, however, a belief about legitimacy cannot exist without a *non*-specific belief that some treatment is appropriate.

[43] Personal moral LR seems to correspond to what Feinberg (1973: 1-2) calls *observantia* respect (cf. Cranor 1983: 109-11, although I disagree with some of Cranor's remarks).

ject of RR taken to deserve some treatment because of a belief about legitimacy or because of a belief about merit?). It's important to note that both LR and MR are kinds of RR, not just of PR; i.e., LR and MR are compatible with NC. This remark helps to explain why NC hovers between respect and contempt: there are two kinds of NC, a kind based on a belief about legitimacy or merit (*negative legitimacy/merit respect*) and a kind based on a belief about illegitimacy or demerit (*illegitimacy/demerit contempt*). If I'm almost certain that I would kill or severely wound you in a duel, I may still believe that you deserve to be challenged (negative treatment, NC) for either of two reasons: I may want to preserve your honor because I view you as a worthy adversary ((negative) merit respect), or I may want to exterminate you because I view you as evil (demerit contempt).[44]

1.2.3. Defiance contempt and disregard contempt

A belief about legitimacy or merit is compatible not only with NC, but also with IC: I may believe that the object's legitimacy or merit makes the object deserve a certain treatment from other people but not from me (*defiance contempt*). Even if I believe that, because you are a nobleman, your challenge to a duel would deserve a response from people who take dueling seriously, I may still believe that your challenge deserves no response from me because I consider dueling silly.[45] Defiance contempt has the paradoxical feature of combining aspects of both respect and contempt.[46] If I recognize you

---

[44] *Positive illegitimacy/demerit respect* can also exist; see footnote 36 for an example.

[45] The belief that dueling is silly is compatible with the belief that a challenge to a duel deserves a response from people who take dueling seriously.

[46] This paradoxical feature of defiance contempt differs from the related paradoxical feature of NC: *every* instance of defiance contempt combines aspects of both respect and contempt, whereas *some* instances of NC are akin to respect and *other* instances of NC are akin to contempt. This difference suggests that the "paradox" in the case of defiance contempt cannot be resolved as it was in the case of NC (where two kinds of NC were distinguished).

as a dangerous opponent but I defy your attack by exhibiting "contempt of danger", then I respect your attack insofar as I admit that it deserves in general to be treated with caution, but I also despise it insofar as I believe that its merit gives *me* no reason to treat it with caution.[47] Defiance contempt can be contrasted with a more usual kind of IC, namely *disregard contempt*, which is based on a belief about nonlegitimacy or nonmerit:[48] I may believe that your challenge to a duel deserves no response from me because I'm a nobleman and you are a servant, so that your challenge is beneath contempt.[49, 50, 51]

1.2.4. Recognition respect versus appraisal respect

The concepts of RR and IC can be further clarified by examining their relationships with the concepts of AR and AC. I will argue first that (personal moral) IC is in conflict with AR and with AC. If I have AR or AC for someone, then I must believe that, at least to some extent, she is a moral agent and accountable for her actions, so that there

---

[47] This example should be distinguished from cases in which I do believe that I have some reason to treat your attack with caution but either I take this reason to be overriden by some contrary reason or I nevertheless choose to behave recklessly; in such cases I can have MR for your attack and not behave accordingly.

[48] I distinguish *non*legitimacy (i.e., lack of legitimacy) from *il*legitimacy. As I use these terms, a judgment of nonlegitimacy is compatible with IC, but a judgment of illegitimacy is not. For example, I would say that my crazy neighbor's innocuous claim to the throne is nonlegitimate and deserves no reaction, whereas the subversive claim to the throne made by the rightful heir's brother is illegitimate and should be actively opposed. (Similarly for *non*merit and *de*merit.)

[49] It seems that the expression "beneath contempt" refers sometimes to IC ("his suggestion is beneath contempt") and other times to extremal AC ("such conduct is beneath contempt").

[50] Some instances of IC are intermediate between defiance and disregard contempt. I may believe that your challenge to a duel deserves no response from me because I'm a prince and you are a mere baron; then I consider your challenge as not having *enough* merit to give *me* a reason to respond to it.

[51] For the sake of simplicity I omitted from Table 1.2 obstacle respect (see footnotes 31 and 37) and a third kind of IC, *nonconscious contempt*. If the thought that slaves deserve to be treated in accordance with basic human rights has never crossed the mind of a slave owner, then she can have nonconscious contempt for slaves. The slave owner can have (e.g.) disregard contempt for slaves only if she consciously rejects the claim that there is some treatment which they deserve.

are circumstances in which she would deserve to be praised or blamed. But we saw that IC for someone presupposes the belief that, as far as her deserts go, it's morally indifferent how she is treated; hence the conflict. Objection: can't a slave owner believe that a particular slave is both subhuman and a bad person? Two points in reply. First, the conflict in question is not strictly speaking an inconsistency. A slave owner who really considers a slave as indifferently as an inanimate piece of property but also sees in the slave's misdemeanors a manifestation of the slave's wickedness may be making the *normative* mistake of denying that the slave's status as a moral agent entitles the slave to certain kinds of treatment (cf. Boxill 1976/1995: 98).[52] Second, I may view someone as subhuman without having IC for her: I may believe that she is a moral agent *inferior* to myself. In such a case I can have *inegalitarian*, more specifically *condescending* MR for her. Condescending MR is a third concept that vacillates between respect and contempt (the previous two being NC and defiance contempt). I will examine a kind of inegalitarian MR in more detail later on (§1.2.5); for the moment I just note that, interestingly, the other kind of inegalitarian MR, namely *deferential* MR (in which I believe that the object of MR has *more* merit than myself), is clearly not a kind of contempt.[53]

I turn now to a comparison between MR and AR. Both MR and AR presuppose a positive evaluation; thus it may be fairly asked how MR differs from AR. First, MR is wider than AR, in the sense that AR, unlike MR, is normally personal and moral.[54] Second, (even personal moral) MR differs from AR because one can have AR without hav-

---

[52] If one accepts that being a moral agent entails having subject dignity (§3.2), then the slave owner may be denying that object dignity (§3.2) always accompanies subject dignity.

[53] But is it a kind of *self*-contempt? (Cf. footnote 9 on deferential AR.)

[54] Non-moral AR does exist, but depends partly on a moral evaluation of the object (footnote 4). Non-personal AR does exist, but its objects are normally (though not always: footnote 3) entities related to persons (e.g., character traits or performances).

ing MR:[55] I can believe that you are a good person without believing, e.g., that you deserve to be praised (maybe because I believe that people deserve to be praised only for supererogatory actions). It may be objected that AR does have a cognitive$_3$ component: a dispositional belief that the object deserves a positive behavioral response. I reply that there are two differences between the cognitive$_3$ components of AR and of MR. First, normally MR is specific whereas AR is not: usually when I have AR for you I don't have in mind any specific behavioral response that you deserve.[56] Second, the cognitive$_3$ component of AR is a *conditional* disposition: *if* I think that the situation calls for awarding a prize and I have AR for you, then I may have the dispositional belief that you deserve the prize. MR, by contrast, presupposes that a specific condition is fulfilled (e.g., I think that the situation calls for awarding a prize), so that the cognitive$_3$ component of MR may be called a *condition-fulfilled* disposition.

1.2.5. Worth respect (WR)

Although one can have AR without having MR, it may be thought that one cannot have MR without also having AR. Not so, however: another (and a most important) difference between AR and (personal moral) MR is that the positive evaluation which constitutes the cognitive$_1$ component of MR is *not* a positive appraisal of a person in terms of her character. Even if I have AC for you because I consider you to be rude, I can still believe that you deserve to be treated courteously; my attitude is MR insofar as this belief is grounded on my taking you to have a certain *value*. But what is this value, given

---

[55] But how can I have AR without having MR, if I cannot have IC when I have AR? MR (or RR) and IC are contraries, not contradictories: see the formalizations in footnote 35.

[56] It may be objected that MR *can* also be non-specific. But although I argued that the concept of nonspecific *recognition* respect is useful (insofar as it marks a contrast with IC), I don't see any usefulness of a concept of non-specific *merit* respect. (By contrast, I claim that the concept of non-specific AR is useful because AR *is* normally non-specific.)

that it is not moral goodness?[57] It could be what has been called *dignity*, understood as *inherent worth*.[58] The idea is that there is something good about persons which is independent of how good (or bad) persons they are.[59] I will call *worth respect* (WR) the kind of moral MR which presupposes a positive evaluation in terms of (inherent) worth.[60] Worth need not be restricted to persons (Hicks 1971; Lombardi 1983; P. W. Taylor 1981): one can distinguish *narrow-scope* from *wide-scope* WR, according to whether only persons or also other entities (e.g., animals or plants) are taken to have worth. One can also distinguish *egalitarian* WR, which presupposes the belief that all entities which have worth have the *same* worth, from *inegalitarian* WR, which presupposes the belief that some entities which have worth have less worth than others. Wide-scope WR can be egalitarian (P. W. Taylor 1981) or inegalitarian (Hicks 1971; Lombardi 1983);[61] similarly for narrow-scope WR, although narrow-scope inegalitarian WR (which presupposes the belief that some persons have less worth than others) seems to be an unpopular view.

---

[57] Why can't this value be moral goodness? One might claim that AR and MR are grounded on the *same* kind of value, AR presupposing an assessment of *high* value and MR presupposing an assessment of *positive* value. (To avoid the objection that MR would then amount to non-extremal AR (§1.1.1), one might claim in addition that every degree of positive value corresponds to the same degree of merit, so that MR (in contrast to non-extremal AR) does not admit of degree.) But then it would be impossible to have both MR and AC for someone.

[58] Besides being understood as inherent worth, dignity can be understood as a personal quality (Kolnai 1976/1995: 60; cf., e.g., talk of "dignified demeanor") or as *status dignity*, which one has in virtue of one's rank or place in a hierarchy (cf., e.g., dignity as pertaining to "dignitaries"; Kolnai 1976/1995: 59). Status dignity has "public availability" (Meyer 1989: 522; cf. Dillon 1995: 23).

[59] I have no definition of worth to offer, save for saying that worth is a kind of *intrinsic* value (P. W. Taylor 1981: 201). Given that I view worth as the ground for desert, I cannot say without circularity that having worth *consists* in deserving a certain treatment (contrast P. W. Taylor 1981: 201). Note that inherent worth, which depends on one's (essential) nature, differs from the worth that one has in virtue of the way in which one conducts one's life (cf. Deigh 1983/1995: 150-1).

[60] Status dignity (footnote 58) can guarantee legitimacy; therefore, by substituting status dignity for inherent worth in the definition of WR, one can get a kind of LR (rather than MR).

[61] Wide-scope inegalitarian WR is compatible with the belief that all *persons* have equal worth: other beings may be taken to have less (or more) worth than persons.

Philosophers who believe that inegalitarian WR is somehow mistaken or inappropriate face the *problem of equal worth* (PEW), namely the problem of explaining why all beings that have worth have the same worth. No matter whether one believes that wide-scope WR is or is not appropriate, one faces the *problem of the ground of worth* (PGW), namely the problem of explaining why some beings have worth whereas others don't.[62] Solving PGW does not automatically solve PEW: if the ground of worth is, e.g., the capacity to act for reasons, then the question remains why slight differences in this capacity result in the same level of worth (if they do). My object here is not to resolve these normative problems; it is rather to point out that inegalitarian WR and wide-scope WR are possible kinds of respect (if one accepts the idea of inherent worth in the first place).

The claim that inegalitarian (in particular, condescending) WR is a possible kind of respect can be contested. Inegalitarian WR admits of degree: if I believe that Ethel has a higher worth than Eunice, then there is a sense in which I can have more WR for Ethel than for Eunice.[63] (This can be the case even if I believe that Ethel and Eunice have equally good character: I may believe that Ethel has more worth because she is of royal blood.) Darwall, however, claims that, "if all persons as such should be treated equally, there can be no degrees of recognition respect for them" (1977/1995: 192). Is Darwall

---

[62] PEW should be distinguished from the *problem of equal respect* (PER), namely the problem of explaining why all beings that have worth deserve the same treatment: two beings can deserve the same treatment even if they have different worth. (Usually the problem of equal respect is formulated only for narrow-scope (or for personal) WR; cf., e.g., Wong 1984.) Similarly, PGW should be distinguished from the *problem of the ground of respect* (PGR), namely the problem of explaining why the possession of worth makes a being deserve a certain treatment.

[63] There are also other senses in which I can have more WR for Ethel than for Eunice. I may believe that Ethel deserves to be treated more positively than Eunice. (This belief doesn't follow from the belief that Ethel has more worth: I may believe that persons of slightly different worth deserve to be treated equally—cf. footnote 62.) Or I may have a stronger disposition to treat Ethel as I think she deserves to be treated. Darwall (1977/1995: 191-2) notes that there is a sense in which even egalitarian personal WR can be said to admit of degree: "one may be a greater or lesser respecter of persons", depending on what kind of treatment one takes worth to make appropriate.

confusing descriptive with normative issues here? One can accept both the normative claim that inegalitarian WR is inappropriate and the descriptive claim that Quentin, who thinks that women have worth but are inferior to men, still has (condescending) WR for women.[64] Darwall might reply that Quentin's wife could justifiably complain that Quentin has no respect for her. It may be more precise, however, to say that Quentin lacks *proper* respect for his wife, not that he has no WR at all.[65] After all (we are supposing), Quentin has no IC or demerit contempt for his wife. Using the terminology that I introduced when I discussed mistakes related to AR (§1.1.4), it seems that Quentin is making a normative rather than a conceptual mistake. Darwall's view, however, has some attractiveness (Hill 1973/1995a), so it may help to consider a second example. Compare a parent who thinks that her children have less worth than fully grown moral agents with a slave owner who thinks that her adult slaves are like children and thus have less worth than fully grown moral agents. It seems that on Darwall's view neither the parent nor the slave owner has WR. But this is clearly implausible in the case of the parent, and it's almost equally implausible in the case of the slave owner, given that the only relevant difference between the slave owner and the parent is that the former, unlike the latter, is making a factual mistake (the slaves *are* fully grown moral agents).[66]

If the issue were purely terminological, then the dispute could be resolved by introducing a new distinction (which there is good reason to introduce anyway—cf. Massey 1983/1995): one could say that Quentin and the slave owner have *subjective* but not *objective* WR. My worry, however, is that Darwall's view seems to "solve" the

---

[64] Cf. Deigh (1983/1995: 146-7): in characterizing an emotion, one should not specify the conditions under which the emotion is experienced rationally, because the fact that one never has good reason to feel an emotion does not in itself show the characterization of the emotion to be faulty.

[65] I have no quarrel with the following modification of Darwall's statement: "if all persons as such should be treated equally, then there are no degrees of *appropriate* (*proper*) recognition respect for them".

[66] Assume that the slave owner believes that slaves, like children, *can* become fully grown moral agents (after several years of "education").

problem of equal worth by definitional fiat: if it's *impossible* to have more (objective) WR for some persons than for others, then the question why one should have the same WR for all persons does not even arise. Darwall might reply that, on his view, the question remains why one should have WR for persons at all. True, but this question pertains to the problem of the *ground* of worth, and as I explained in the last paragraph but one, solving PGW does not automatically solve PEW. To put the point differently: Darwall's view seems to presuppose a negative answer to the normative question of whether inegalitarian WR is ever justified. It may be objected that I have too rigid a conception of the descriptive/normative distinction: the intuition that inegalitarian personal WR is never justified is so deep (the objection goes) that it must function as a constraint for the descriptive task. But even if I agree that the descriptive/normative distinction must be taken with a grain of salt, and even if I share the intuition that inegalitarian WR is never justified, I still want to leave open the possibility that at the end of my inquiry I will give up this intuition. (Actually, the example of the parent in the previous paragraph already indicates that inegalitarian personal WR may be sometimes justified.)[67]

2. Kant on respect

Having completed my elaboration of Darwall's (1977/1995) distinction between AR and RR, I will argue now that a similar distinction is to be found in Kant's writings on *Achtung* ('respect'). I will first explain (§2.1) that Kant identifies three main kinds of *Achtung*. Then I will examine these three kinds successively: the feeling of respect for

---

[67] Darwall (personal communication) claims that he intended to deny only that there can be degrees of RR for a person *as such* (i.e., RR for a person purely on the ground that she is a person). Now as a formal point there can be indeed no degrees of RR for persons as persons, but there can be no degrees of RR for persons as (e.g.) living beings either. But why should we be interested in RR for persons as persons? We don't know in advance whether this kind of RR captures the idea of *dignity* that WR is intended to capture. (Except if a person is defined as a being that has dignity; but as Frankena (1986: 152) points out, this *loaded* concept of a person does not enable us to find out which beings are persons.)

the moral law (§2.2), the feeling of appraisal respect for a person (§2.3), and the maxim of engaging in recognition-respect-appearing behavior (§2.4).[68]

2.1. Three kinds of Achtung

The fundamental kind of *Achtung* for Kant is respect for the moral law (*Achtung fürs Gesetz*[69]), which Kant understands as a feeling (*Gefühl*). But it's a feeling of a special kind: "it is not a feeling *received* through outside influence, but one *self-produced* by a rational concept".[70] This feeling is the effect (*Wirkung*) that (awareness of) the moral law has on a person,[71] and is the only feeling "which we can know completely a priori".[72] Respect for the moral law "is identical with consciousness of one's duty", and "in its subjective aspect is called moral feeling"[73] (*moralisches Gefühl*), which Kant under-

---

[68] Here is an example of the way in which I refer to Kant's works throughout this chapter: G4.401.19-21/T69n/A16n. 'G' stands for *Groundwork of the metaphysic of morals*; in the place of 'G' can be 'P' (for *Critique of practical reason*) or 'M' (for *The metaphysics of morals*). '4.401.19-21' refers to volume 4 (5 for P, 6 for M), page 401, lines 19-21 of the Prussian Academy edition of Kant's works. 'T69n' refers to page 69 (footnote) of the translation I used (see list of references at the end of the dissertation). 'A16n' refers to page 16 (footnote) of the first German edition—only for G (1785 edition) and P (1788 edition). I use modernized German spelling throughout (e.g., 'Wert' for Kant's 'Werth'). Sometimes I use my own translation.

[69] Or: "Achtung fürs moralische Gesetz". In *The metaphysics of morals* Kant also uses the expression "Achtung vor dem Gesetz(e)" (M6.403.05,6.410.25,6.464.05/T162,168,210).

[70] "Allein wenn Achtung ein Gefühl ist, so ist es doch kein durch Einfluß empfangenes, sondern durch einen Vernunftbegriff selbstgewirktes Gefühl" (G4.401.19-21/T69n/A16n).

[71] "Die unmittelbare Bestimmung des Willens durchs Gesetz und das Bewußtsein derselben heißt Achtung, so daß diese als Wirkung des Gesetzes aufs Subjekt und nicht als Ursache desselben angesehen wird" (G4.401.25-28/T69n/A16n).

[72] "Also ist Achtung fürs moralische Gesetz ein Gefühl, welches durch einen intellektuellen Grund gewirkt wird, und dieses Gefühl ist das einzige, welches wir völlig a priori erkennen, und dessen Notwendigkeit wir einsehen können" (P5.073.34-37/T77/A73).

[73] "Die Achtung vor dem Gesetze, welche subjektiv als moralisches Gefühl bezeichnet wird, ist mit dem Bewußtsein seiner Pflicht einerlei" (M6.464.05-06/T210).

stands as "the susceptibility to feel pleasure or displeasure merely from being aware that our actions are consistent with or contrary to the law of duty."[74]

A second kind of *Achtung* that Kant identifies is (appraisal) respect for a person (*Achtung für eine Person*), for which he sometimes uses (in *The metaphysics of morals*) the Latin *reverentia*, and which he also understands as a feeling. Respect for a person is what I have called the feeling of self-comparative deferential AR (footnote 9). It's the feeling "that comes from comparing our own *value* with another's",[75] value (*Wert*) being moral goodness:[76] respect is a "tribute" we pay to "merit"[77] (*Verdienst*). It's the feeling that I experience when I consider "a humble plain man, in whom I perceive righteousness in a higher degree than I am conscious of in myself". I have this feeling because this man's "example holds a law before me which strikes down my self-conceit when I compare my own conduct with it".[78] Therefore, according to Kant, respect for a person is a

---

[74] "[Das moralische Gefühl] ist die Empfänglichkeit für Lust oder Unlust bloß aus dem Bewußtsein der Übereinstimmung oder des Widerstreits unserer Handlung mit dem Pflichtgesetze" (M6.399.19-21/T160).

[75] "[D]as Gefühl aus der Vergleichung unseres eigenen Werts mit dem des anderen" (M6.449.24-25/T199). I have substituted "value" for "worth" in Gregor's translation.

[76] The examples that Kant goes on to give are not of moral goodness: "such as a child feels merely from habit towards his parents, a pupil toward his teacher, or any subordinate towards his superior" (M6.449.25-27/T199). But this is because in the context of the specific passage Kant is not concerned with explaining what kind of value is relevant to respect for a person. (He is only juxtaposing feelings of respect with respect-related behavior, his third kind of *Achtung*.) Thus my claim that the relevant kind of value is moral goodness is based on *other* passages (see footnotes 77 and 78).

[77] "Achtung ist ein Tribut, den wir dem Verdienste nicht verweigern können, wir mögen wollen oder nicht; wir mögen allenfalls äußerlich damit zurückhalten, so können wir doch nicht verhüten, sie innerlich zu empfinden" (P5.077.15-18/T80/A137).

[78] "[V]or einem niedrigen, bürgerlich-gemeinen Mann, an dem ich eine Rechtschaffenheit des Charakters in einem gewissen Maße, als ich mir von mir selbst nicht bewußt bin, wahrnehme, bückt sich mein Geist, ich mag wollen oder nicht [...]. Warum das? Sein Beispiel hält mir ein Gesetz vor, das meinen Eigendünkel niederschlägt, wenn ich es mit meinem Verhalten vergleiche, und dessen Befolgung, mithin die Tunlichkeit desselben, ich durch die Tat bewiesen vor mir sehe" (P5.077.01-04,06-09/T80/A136).

derivative kind of *Achtung*: "All respect for a person is properly only respect for the law (of righteousness and so on) of which that person gives an example."[79]

In *The metaphysics of morals*, Kant explicitly distinguishes feelings of respect from a third kind of *Achtung*, which is "to be understood as the *maxim* of limiting our self-esteem by the dignity of humanity in another person, and so as respect in the practical sense (*observantia aliis praestanda*)".[80] It's only in this sense of *Achtung* that one can talk about a duty of respect: there can be no duty to have feelings of respect. This is made clear by Kant's comparison of love with respect. "*Love* is a matter of *feeling*, not of willing, and I cannot love because I *will* to, still less because I *ought* to (I cannot be constrained to love); so a *duty to love* is an absurdity"[81] (*Unding*). "Respect (*reverentia*) is, again, something merely subjective, a feeling of a special kind, not a judgment about an object that it would be a duty to bring about or promote."[82] But the distinction be-

---

[79] "Alle Achtung für eine Person ist eigentlich nur Achtung fürs Gesetz (der Rechtschaffenheit etc.), wovon jene uns das Beispiel gibt" (G4.401.35-36/T69n/A17n). I have substituted (here and elsewhere) "respect" for "reverence" and "righteousness" for "honesty" in Paton's translation. Cf.: "Because we regard the development of our talents as a duty, we see too in a man of talent a sort of *example of the law* (the law of becoming like him by practice), and this is what constitutes our respect for him" ("Weil wir die Erweiterung unserer Talente auch als Pflicht ansehen, so stellen wir uns an einer Person von Talenten auch gleichsam das Beispiel eines Gesetzes vor (*ihr durch Übung hierin ähnlich zu werden*) und das macht unsere Achtung aus": G4.401.37-39/T69n/A17n; the parenthetical remark was added in the second (1786) edition). Cf. also: "If one examines more accurately the concept of respect for persons, as this has been previously presented, one will perceive that it always rests on the consciousness of a duty which an example holds before us" ("Wenn man den Begriff der Achtung für Personen, so wie er vorher dargelegt worden, genau erwägt, so wird man gewahr, daß sie immer auf dem Bewußtsein einer Pflicht beruhe, die uns ein Beispiel vorhält": P5.081.30-32/T85n/A144n-145n).

[80] "[...] sondern nur eine Maxime der Einschränkung unserer Selbstschätzung durch die Würde der Menschheit in eines anderen Person, mithin die Achtung im praktischen Sinne (*observantia aliis praestanda*) verstanden wird" (M6.449.27-30/T199).

[81] "Liebe ist eine Sache der Empfindung, nicht des Wollens, und ich kann nicht lieben, weil ich will, noch weniger aber, weil ich soll (zur Liebe genötigt werden); mithin ist eine Pflicht zu lieben ein Unding" (M6.401.24-26/T161). Note that Kant does talk about a duty to love, but only when he understands love as a maxim (which he contrasts to love as a feeling: M6.449.16-22/T199; cf. G4.399.28-34/T67/A13).

[82] "Achtung (*reverentia*) ist ebensowohl etwas bloß Subjektives; ein Gefühl eigener Art, nicht ein Urteil über einen Gegenstand, den zu bewirken oder zu befördern, es eine Pflicht gäbe" (M6.402.29-31/T162). What Kant goes on to say may suggest that his reason for claiming that there can be no duty to have feel-

tween *Achtung* as *reverentia* and *Achtung* as *observantia* is not merely the distinction between feelings of respect and respect-related behavior (or duties). After all, according to Kant, there *is* a feeling of respect that corresponds to *observantia*: "*Love* and *respect* are the feelings that accompany the carrying out of those duties [of love and respect]".[83] Clearly, the feeling that accompanies the carrying out of the duty of respect is distinct from the feeling that one has when one considers a righteous person. Thus I suggest that the distinction between *reverentia* and *observantia* also corresponds to the distinction between appraisal and recognition respect. This connection between Kant's corpus and recent philosophical discussion on respect seems to have escaped commentators' notice; indeed, it seems that only few commentators (e.g.: Feinberg 1973; Gregor 1963: 181; Massey 1983: 60-4; Paton 1948: 65 n. 2) are even aware of Kant's *reverentia/observantia* distinction.[84]

I proceed now to examine each of the above three kinds of *Achtung* in more detail.[85]

---

ings of respect is not an analogy between love and respect, but rather a regress argument: "Denn sie könnte, als Pflicht betrachtet, nur durch die Achtung, die wir vor ihr haben, vorgestellt werden. Zu dieser also eine Pflicht zu haben würde soviel sagen, als zur Pflicht verpflichtet werden" (M6.402.31-34/T162). I think it's better to regard this argument as an *additional* reason (rather than as Kant's *only* reason) for the claim that there can be no duty to have feelings of respect; be that as it may, my only concern here is to point out that Kant makes this claim.

[83] "Liebe und Achtung sind die Gefühle, welche die Ausübung dieser Pflichten begleiten" (M6.448.14-15/T198).

[84] Cranor (probably misunderstanding Paton 1948: 65 n. 2) claims that "[i]n his later works, including *The Metaphysical Principles of Virtue* [..., Kant] translated *Achtung* as the Latin *observantia*" (1980: 20). But Kant is careful to translate *Achtung* sometimes as *reverentia* and sometimes as *observantia* in *The Metaphysical Principles of Virtue.*

[85] Note that *Respekt* ('awe'; in Kant's spelling: *Respect*) is relevantly used only once in Kant's corpus (M6.438.14/T189) and is equated by Kant with "respect coupled with fear" ("mit Furcht verbundene Achtung"). Cf.: Paton 1948: 64; Feinberg 1973; Massey 1983: 60.

2.2. Achtung *as respect for the moral law*

Why does Kant claim that respect for the moral law is a feeling "which we can know completely a priori and the necessity of which we can discern"?[72] Kant's starting point is the observation that some actions occur for the sake of the moral law: the law determines the will directly, not by means of any feeling, no matter of what kind, which must be presupposed if the law is to become a sufficient determinant of the will.[86] But how can a law directly determine the will? Kant has no answer to propose: he claims that this is "an insoluble problem for the human reason", and "identical with the problem of how a free will is possible."[87] But we know that it happens; we know that sometimes the moral law *is* a *drive* (*Triebfeder; elater animi*), a subjective determinant of the will.[88] Therefore, we can investigate a priori what effect the moral law has (better put: *must* have) on the mind (*Gemüt*); and this Kant sets out to do.[89]

Kant's fundamental observation is that, whenever an action occurs for the sake of the moral law but the agent has inclinations (*Neigungen*) contrary to the law, the law

---

[86] "Das wesentliche alles sittlichen Werts der Handlungen kommt darauf an, daß das moralische Gesetz unmittelbar den Willen bestimme. Geschieht die Willensbestimmung zwar gemäß dem moralischen Gesetze, aber nur vermittelst eines Gefühls, welcher Art es auch sei, das vorausgesetzt werden muß, damit jenes ein hinreichender Bestimmungsgrund des Willens werde, mithin nicht um des Gesetzes willen: so wird die Handlung zwar Legalität, aber nicht Moralität enthalten" (P5.071.28-34/T75/A126-7; cf. M6.214.13-19/T14).

[87] "Denn wie ein Gesetz für sich und unmittelbar Bestimmungsgrund des Willens sein könne (welches doch das Wesentliche aller Moralität ist), das ist ein für die menschliche Vernunft unauflösliches Problem und mit dem einerlei: wie ein freier Wille möglich sei" (P5.072.21-24/T75/A128).

[88] "Wenn nun unter Triebfeder (*elater animi*) der subjektive Bestimmungsgrund des Willens eines Wesens verstanden wird, dessen Vernunft nicht, schon vermöge seiner Natur, dem objektiven Gesetze notwendig gemäß ist, [...]" (P5.071.34-072.02/T75/A127). Actually Kant should be using *Willkür*, not *Wille* (see footnote 92); cf.: "eine Triebfeder, welche den Bestimmungsgrund der Willkür [...]" (M6.218.15-16/T20).

[89] "Also werden wir nicht den Grund, woher das moralische Gesetz in sich eine Triebfeder abgebe, sondern was, so fern es eine solche ist, sie im Gemüte wirkt (besser zu sagen, wirken muß), a priori anzuzeigen haben" (P5.072.24-27/T75-6/A128).

thwarts (*Abbruch tut*) these inclinations.[90] Some of these inclinations constitute the agent's *self-conceit* (*Eigendünkel; arrogantia*), namely the tendency to view oneself as a worthy (important) person,[91] the tendency to view the subjective determinants of one's choice (*Willkür*)[92] legislatively, as unconditional practical principles.[93] In contrast to *self-love* (inclinations to promote one's self-interest),[94] which the moral law thwarts but does not annihilate (because one can promote one's self-interest within the limits imposed by the moral law), self-conceit is annihilated, struck down (*niedergeschlagen*) by the moral law, because "all claims of self-esteem which precede conformity to the moral law are null and void".[95] But whatever, in our own judgment, thwarts our self-love or strikes

---

[90] "Das Wesentliche aller Bestimmung des Willens durchs sittliche Gesetz ist: daß er als freier Wille, mithin nicht bloß ohne Mitwirkung sinnlicher Antriebe, sondern selbst mit Abweisung aller derselben, und mit Abbruch aller Neigungen, so fern sie jenem Gesetze zuwider sein könnten, bloß durchs Gesetz bestimmt werde" (P5.072.28-32/T76/A128).

[91] Kant does not explicitly define *Eingendünkel* in this way, but I infer this definition from what Kant says in P5.073.09-27/T76/A129-30, especially from his apparent tendency to use *Eigendünkel* and *Selbstschätzung* ('self-esteem') interchangeably. Cf.: "But lack of modesty in one's claims to be **respected** by others is self-conceit" ("Die Unbescheidenheit der Forderung aber, von anderen geachtet zu werden, ist der Eigendünkel": M6.462.09-10/T209).

[92] For the distinction between will (*Wille*) and choice (*Willkür*), see M6.213.14-26/T13 and Cranor (1980: 23). As Lauener (1981: 250-2) points out, in the *Critique of practical reason* Kant does not always carefully distinguish between will and choice.

[93] "Man kann diesen Hang, sich selbst nach den subjektiven Bestimmungsgründen seiner Willkür zum objektiven Bestimmungsgrunde des Willens überhaupt zu machen, die Selbstliebe nennen, welche, wenn sie sich gesetzgebend und zum unbedingten praktischen Prinzip macht, Eigendünkel heißen kann" (P5.074.15-19/T77/A131). Given that this passage is almost contiguous to the first passage referred to in footnote 91, Kant probably took the two definitions of *Eigendünkel* to be equivalent.

[94] "This [self-regard] consists either of self-love, which is a predominant benevolence toward one's self (*philautia*) or of self-satisfaction (*arrogantia*). The former is called, more particularly, selfishness; the latter, self-conceit" ("Diese [die Selbstsucht] ist entweder die der Selbstliebe, eines über alles gehenden Wohlwollens gegen sich selbst (philautia), oder die des Wohlgefallens an sich selbst (arrogantia). Jene heißt besonders Eigenliebe, diese Eigendünkel": P5.073.11-14/T76/A129).

[95] "Die reine praktische Vernunft tut der Eigenliebe bloß Abbruch, indem sie solche, als natürlich, und noch vor dem moralischen Gesetze, in uns rege, nur auf die Bedingung der Einstimmung mit diesem Gesetze einschränkt [...]. Aber den Eigendünkel schlägt sie gar nieder, indem alle Ansprüche der Selbstschätzung, die vor der Übereinstimmung mit dem sittlichen Gesetze vorhergehen, nichtig und ohne Befugnis sind" (P5.073.14-21/T76/A129-30).

down our self-conceit, humiliates (*demütigt*) us, and thus awakens respect for itself insofar as it is a positive determinant of the will.[96] This is why Kant claims that respect for the moral law is a feeling which we can know a priori: we can know a priori that such a feeling must exist,[97] even though we are unable to describe the mechanism by means of which it is produced.

Is Kant's view self-consistent? On the one hand, Kant claims that sometimes the moral law determines the will directly, not by means of any feeling, *no matter of what kind.*[86] On the other hand, Kant posits a feeling of a special kind as a necessary concomitant of moral action. Did Kant "compromise the effort" (Nagel 1970: 11) to establish reason as a sufficient source of moral motivation? Or did he unwittingly admit that "some motivational factor [is] necessary to account for being moved to act on moral grounds, beyond the mere recognition of the law" (Bond 1983: 11)? I don't think that Kant is guilty of a lapse here. Kant does not claim that the feeling of respect must be *presupposed* if the law is to become a sufficient determinant of the will.[86] Nor does he claim that respect for the law exerts some motivational influence on the will. In fact, he claims the opposite: "respect for the law is not the drive to morality; it is morality itself, re-

---

[96] "Was nun unserem Eigendünkel in unserem eigenen Urteil Abbruch tut, das demütigt. [...] Dasjenige, dessen Vorstellung, als Bestimmungsgrund unseres Willens, uns in unserem Selbstbewußtsein demütigt, erweckt, so fern als es positiv und Bestimmungsgrund ist, für sich Achtung" (P5.074.23-24,26-29/T77-8/A132). Cf.: "Since this law, however, is in itself positive, [...] it is at the same time an object of respect, since [...] it weakens self-conceit. And as striking down, i.e., humiliating, self-conceit, it is an object of the greatest respect and thus the ground of a positive feeling, which is not of empirical origin" ("Da dieses Gesetz aber doch etwas an sich Positives ist, [...] so ist es, indem es [...] den Eigendünkel schwächt, zugleich ein Gegenstand der Achtung, und, indem es ihn sogar niederschlägt, d. i. demütigt, ein Gegenstand der größten Achtung, mithin auch der Grund eines positiven Gefühls, das nicht empirischen Ursprungs ist": P5.073.27-34/T77/A130).

[97] Assuming the a posteriori claim that some actions occur for the sake of the moral law (cf. footnote 89: "so fern es eine solche ist [...]"). See Broadie & Pybus (1975: 61-2) for related views on what Kant may mean when he says that respect is a feeling known a priori.

garded subjectively as a drive".[98] In Sytsma's (1993: 121) words: "The feeling of respect is [...] the *epiphenomenon* of moral motivation."

Commentators have been puzzled about the way in which the moral law could directly determine the will. As Reath points out, "Kant does not think that the Moral Law determines the will through a quasi-mechanical or affective force"; "Kant's conception of choice should not be understood on the analogy of a sum of vector forces (or of mechanical forces acting on an object)". But then "it becomes harder to see how [the Moral Law] can counteract inclinations, though Kant surely thinks that it does" (1989: 290, 291). Disregarding Kant's explicit and repeated assertions to the contrary,[99] Timmons (1985: 391) attributes to Kant a view according to which "consciousness of the moral law [...] can only be said to affect the will *indirectly*"! Otherwise, Timmons claims, Kant's view involves "a certain mysterious element". But Kant was the first to admit that the phenomenon of moral motivation is "wholly impossible to comprehend";[100] at least, Kant claimed, "we do comprehend its *incomprehensibility*".[101]

---

[98] "Und so ist die Achtung fürs Gesetz nicht Triebfeder zur Sittlichkeit, sondern ist die Sittlichkeit selbst, subjektiv als Triebfeder betrachtet" (P5.076.04-06/T79/A134). This passage indicates how other, seemingly contradictory passages are to be understood. E.g.: "Respect for the moral law is therefore the sole and undoubted moral drive" ("Achtung fürs moralische Gesetz ist also die einzige und zugleich unbezweifelte moralische Triebfeder": P5.078.20-21/T82/A139).

[99] "Die Vernunft bestimmt in einem praktischen Gesetze unmittelbar den Willen, nicht vermittelst eines dazwischen kommenden Gefühls der Lust und Unlust, selbst nicht an diesem Gesetze" (P5.025.06-08/T24/A45). Cf.: P5.046.35-36/T48/A81; P5.048.10-11/T49/A83.

[100] "Es ist aber gänzlich unmöglich, einzusehen, d. i. a priori begreiflich zu machen, wie ein bloßer Gedanke, der selbst nichts Sinnliches in sich enthält, eine Empfindung der Lust oder Unlust hervorbringe" (G4.460.12-15/T128/A123). Cf. footnote 87.

[101] "Und so begreifen wir zwar nicht die praktische unbedingte Notwendigkeit des moralischen Imperativs, wir begreifen aber doch seine Unbegreiflichkeit" (G4.463.29-31/T131/A128). The question of whether moral motivation is indeed mysterious is controversial and too large to be discussed here (cf. Vranas 1995).

2.3. Achtung *as* reverentia*: the feeling of appraisal respect for persons*

When I consider a morally good (e.g., a righteous) person, normally I feel AR for that person. Why? Recall that, according to Kant, whatever in my own judgment thwarts my self-love or strikes down my self-conceit humiliates me and awakens thus a feeling of respect in me.[96] But the righteous person's "example holds a law before me which strikes down my self-conceit when I compare my own conduct with it; that it is a law which can be obeyed, and consequently is one that can actually be put into practice, is proved to my eyes by the act".[78] This account of what goes on when I feel AR indicates that, for Kant, the feeling of respect for the moral law and the feeling of AR are manifestations of the same phenomenon: the humiliation that one experiences when something, in one's own judgment, thwarts one's self-love or strikes down one's self-conceit.[102] Nevertheless, the former feeling is primary and the latter derivative, in the sense that what humiliates me when I feel AR is consciousness of the moral law. This explains Kant's claim that "[a]ll respect for a person is properly only respect for the law (of righteousness and so on) of which that person gives an example."[79]

In apparent contradiction to Kant's claim, Velleman interprets Kant as holding that "all reverence for the law is properly only reverence for the person" (1999: 348 n. 30). To understand what Velleman means, note that Kant applies the phrase "object of respect" (*Gegenstand der Achtung*) to various entities, including (Broadie & Pybus 1975: 59-60): the moral law, legislation[103] (*Gesetzgebung*),[104] and our "ideal will" (*Wille in der*

---

[102] Kant's distinction between the way in which the moral law *thwarts* self-love and the way in which the moral law *strikes down* self-conceit (footnote 95) is a distinction of kind, not of degree: self-conceit cannot be thwarted in the way in which self-love can. (Kant sometimes speaks as if the distinction were of degree, but I think he has in mind degrees of humiliation; cf. footnote 96.) I take Kant's general notion of respect to include both the thwarting of self-love and the striking down of self-conceit, although Kant sometimes defines *Achtung* by referring to only one of these two phenomena (cf.: "Eigentlich ist Achtung die Vorstellung von einem Werte, der meiner Selbstliebe Abbruch tut": G4.401.28-29/T69n/A16n).

[103] *Gesetzgebung* is variously translated as "legislation", "law-making", or "enactment of universal law" (cf. Broadie & Pybus 1975: 60).

Idee).[105] Broadie and Pybus argue that these entities "are conceptually so closely related that entitlement to regard any one of them as an object of respect carries with it entitlement to regard each of the others also as an object of respect" (1975: 60). Velleman (personal communication, 11 September 1997) (1) agrees, although he argues that, (2) strictly speaking, the proper object of respect is the ideal will. Velleman uses "the person" synonymously with "the ideal will", so that he understands the claim that "all reverence for the law is properly only reverence for the person" as following from (1) and (2), and as not being about *reverentia* at all. Thus the contradiction is only apparent.[106]

Kant claims that I feel respect when I consider "a humble plain man, in whom I perceive righteousness *in a higher degree* than I am conscious of in myself"[78] (emphasis added). It seems thus that Kant understands *reverentia* as the feeling of *self-comparative deferential* AR[107] (cf. footnote 9). This interpretation is reinforced by the following passage: "I may even be conscious of a like degree of righteousness in myself, and yet re-

---

[104] "[...] it cannot fit as a principle into a possible enactment of universal law. For such an enactment reason compels my immediate respect" ("[...] sie nicht als Prinzip in eine mögliche allgemeine Gesetzgebung passen kann, für diese aber zwingt mir die Vernunft unmittelbare Achtung ab": G4.403.24-26/T70/A20).

[105] "Our own will, provided it were to act only under the condition of being able to make universal law by means of its maxims—this ideal will which can be ours is the proper object of respect" ("Unser eigener Wille, so fern er, nur unter der Bedingung einer durch seine Maximen möglichen allgemeinen Gesetzgebung, handeln würde, dieser uns mögliche Wille in der Idee, ist der eigentliche Gegenstand der Achtung": G4.440.07-10/T107/A86-7).

[106] Velleman takes Kant's claim that "respect for a person is properly only respect for the law" as (1) meant "to rule out persons as proper objects of reverence *insofar as they are inhabitants of the empirical world*", and as (2) compatible with persons' "serving as objects of reverence in their purely intelligible aspect, as instances of rational nature" (1999: 348 n. 30). I agree with (2) but I disagree with (1). Contrary to (1), and as I explained in the last paragraph of the text, Kant's claim under consideration (see also the other passages quoted in footnote 79) relates two kinds of *Achtung*, respect for the moral law and the feeling of AR, and points out that the former is in a sense primary. (One could construe the passage surrounding and including the one quoted in footnote 105 as meant to rule out persons as proper objects of reverence insofar as they are inhabitants of the empirical world, but such an interpretation seems implausible given, among other things, the crucial passage of the "humble plain man" (footnotes 78 and 108). Given that Velleman agrees with my interpretation of Kant's claim, I don't see why Velleman holds (1).

[107] Strictly speaking, *reverentia* involves something more than the feeling of self-comparative deferential AR, namely some kind of focus on one's own moral shortcomings.

spect remains [...], since the man whom I see before me provides me with a standard by appearing to me in a more favorable light in spite of his imperfections, which, though perhaps always with him, are not so well known to me as are my own."[108] I take Kant here to be making the subtle point that I earlier expressed by saying that "the cognitive component of the feeling of AR can be what may be called an *apparent* (as opposed to a real) occurrent belief" (p. 6). But Kant is also implicitly excluding the possibility of feeling respect for a person whom one perceives as righteous but as less righteous than oneself,[109] and I find this view problematic on at least three counts. First, it seems false: I can feel AR for you even if I view you as less good than I am. Second, it seems to contradict Kant's own general account of respect: a person whom I view as less good than I am can still on particular occasions give me an example of the moral law and thus set in motion the humiliation process.[110] Third, it has the unfortunate consequence that I cannot feel *self*-respect,[111] since I cannot view myself as better than myself! I think thus that we had better not take Kant on his word: we should not understand *reverentia* as essentially deferential.[112]

---

[108] "Nun mag ich mir sogar eines gleichen Grades der Rechtschaffenheit bewußt sein, und die Achtung bleibt doch. Denn, da beim Menschen immer alles Gute mangelhaft ist, so schlägt das Gesetz, durch ein Beispiel anschaulich gemacht, doch immer meinen Stolz nieder, wozu der Mann, den ich vor mir sehe, dessen Unlauterkeit, die ihm immer noch anhängen mag, mir nicht so, wie mir die meinige, bekannt ist, der mir also in reinerem Lichte erscheint, einen Maßstab abgibt" (P5.077.09-15/T80/A136-7).

[109] See footnotes 75 and 76 for further support of the claim that Kant's general conception of respect for persons is essentially deferential.

[110] It might be pointed out that the process can be set in motion on particular occasions even when I consider a person whom I view as morally bad. I agree, but I take this point to supplement my critique of *global* AR (Chapter II).

[111] Kant does speak of *reverentia* for oneself: "... und eine unverlierbare Würde (*dignitas interna*) besitzt, die ihm Achtung (*reverentia*) gegen sich selbst einflößt" (M6.436.11-13/T187).

[112] Another claim of Kant's about *reverentia* that I find problematic is the following: "Respect is a tribute we cannot refuse to pay to merit whether we will or not; we can indeed outwardly withhold it, but we cannot help feeling it inwardly" (footnote 77). Kant is ignoring cases in which I have a real but not an apparent belief that one is a good person; cf. the case of "Anne" in Dillon's discussion of basal self-respect (§3.4).

2.4. Achtung *as* observantia*: recognition-respect-appearing behavior*

"The respect that I have for others or that another can require from me (*observantia aliis praestanda*) is [the] recognition of a *dignity* (*dignitas*) in other human beings, that is, of a worth that has no price, no equivalent for which the object evaluated (*aestimii*) could be exchanged."[113] "Humanity itself is a dignity; for a human being cannot be used merely as a means by any human being (either by others or even by himself) but must always be used at the same time as an end. It is just in this that his dignity (personality) consists".[114] Therefore: "The duty of respect for my neighbor is contained in the maxim not to degrade any other to a mere means to my ends (not to demand that another throw himself away in order to slave for my end)."[115]

How broad is the duty of respect? Clearly, it's not the whole of morality, given that Kant contrasts it with the duty of love[115] (cf. footnote 38). But Cranor thinks it's very narrow indeed: "the duty of respect only requires one to refrain from pride, calumny, mockery, not externally manifesting how little one thinks of another, and not censuring a person in the logical use of his reason" (1980: 30). It seems, however, that Cranor (like Hill 1971: 61) misunderstands the list of *examples* that Kant mentions as con-

---

[113] "Achtung, die ich für andere trage, oder die ein anderer von mir fordern kann (*observantia aliis praestanda*), ist also die Anerkennung einer Würde (*dignitas*) an anderen Menschen, d. i. eines Werts, der keinen Preis hat, kein Äquivalent, wogegen das Objekt der Wertschätzung (*aestimii*) ausgetauscht werden könnte" (M6.462.10-15/T209). Cf.: "In the kingdom of ends everything has either a *price* or a *dignity*. If it has a price, something else can be put in its place as *equivalent*; if it is exalted above all price and so admits of no equivalent, then it has a dignity" ("Im Reiche der Zwecke hat alles entweder einen Preis, oder eine Würde. Was einen Preis hat, an dessen Stelle kann auch etwas anderes, als Äquivalent, gesetzt werden; was dagegen über allen Preis erhaben ist, mithin kein Äquivalent verstattet, das hat eine Würde": G4.434.31-34/T102/A77).

[114] "Die Menschheit selbst ist eine Würde; denn der Mensch kann von keinem Menschen (weder von anderen noch sogar von sich selbst) bloß als Mittel, sondern muß jederzeit zugleich als Zweck gebraucht werden, and darin besteht eben seine Würde (die Persönlichkeit)" (M6.462.21-24/T209).

[115] "Die Pflicht der Nächstenliebe kann also auch so ausgedrückt werden: sie ist die Pflicht, anderer ihre Zwecke (sofern diese nur nicht unsittlich sind) zu den meinen zu machen; die Pflicht der Achtung meines Nächsten ist in der Maxime enthalten, keinen anderen Menschen bloß als Mittel zu meinen Zwecken abzuwürdigen (nicht zu verlangen, der andere solle sich selbst wegwerfen, um meinem Zwecke zu frönen)" (M6.450.03-08/T199).

sequences of the duty of respect for an *exhaustive* list of such consequences. There are many more ways than Cranor lists of degrading someone to a mere means to my ends;[116] given the quotation at the end of the last paragraph, any such way runs contrary to the duty of respect.[117]

Kant's statement of the duty of respect[115] is very similar to his second formulation of the Categorical Imperative, the formula of Humanity: "*Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end.*"[118] Indeed, the two statements are sometimes at least implicitly taken to be equivalent (Brink 1997: 268), and commentators often treat the formula of Humanity as a principle of respect for persons (Donagan 1977: 65; Downie & Telfer 1970: 13-5; Frankena 1986: 151; Paton 1948: 165-79; Rawls 1971: 169-70; cf. Cranor 1980: 34 n. 2). But if the two statements are indeed equivalent, contradiction looms: in the *Groundwork* Kant takes the formula of Humanity to entail that one should "further the ends of others",[119] whereas in *The metaphysics of morals* Kant insists that only the duty of love, not the duty of respect, can "be expressed as the duty to make others' ends my own"[115] (cf. Cranor 1980: 30-1).

---

[116] This is not to deny that some of the ways Cranor lists are important; cf. Buss 1999a.

[117] Kant does say that the duty of respect is narrow (*eng*), but only *in comparison* with the duty of love: "Auch wird die Pflicht der freien Achtung gegen andere, weil sie eigentlich nur negativ ist (sich nicht über andere zu erheben), und so der Rechtspflicht, niemandem das Seine zu schmälern, analog, obgleich als bloße Tugendpflicht verhältnisweise gegen die Liebespflicht für enge, die letztere also als weite Pflicht angesehen" (M6.449.31-450.02/T199). (Kant takes ethical duties to be wide and duties of right to be narrow: M6.390.02-03/T153.)

[118] "Handle so, daß du die Menschheit, sowohl in deiner Person, als in der Person eines jeden anderen, jederzeit zugleich als Zweck, niemals bloß als Mittel brauchest" (G4.429.10-12/T96/A66-7). Cf.: P5.087.26-27/T91/A155-6; P5.131.20-25/T138/A237.

[119] "Now humanity could no doubt subsist if everybody contributed nothing to the happiness of others but at the same time refrained from deliberately impairing their happiness. This is, however, merely to agree negatively and not positively with *humanity as an end in itself* unless every one endeavours also, so far as in him lies, to further the ends of others. For the ends of a subject who is an end in himself must, if this conception is to have its *full* effect on me, be also, as far as possible, *my* ends" (G4.430.19-27/T98/A69).

A closer reading, however, reveals that the two statements are *not* equivalent: the formula of Humanity entails but is not entailed by the statement of the duty of respect. This is because the formula of Humanity entails a *conjunctive* injunction (never treat others simply as means, *and* always treat others at least partly as ends), whereas the duty of respect amounts only to the first conjunct of this injunction (never treat others simply as means);[120] Kant derives other-regarding duties from the second conjunct. This is not a desperate attempt to save Kant from contradiction: on the contrary, Kant explicitly claims that "duties to one's fellow human beings arising from the respect due them are expressed only negatively"[121] and admits that "merely to agree negatively and not positively with *humanity as an end in itself*' does not suffice for his derivation of other-regarding duties from the formula of Humanity.[119]

3. Self-respect

To the basic distinction between appraisal and recognition respect corresponds the distinction between *appraisal self-respect* (ASR), whose central component is the belief that one is a good person, and *recognition self-respect* (RSR), whose central component is the belief that one deserves a certain treatment.[122] In general, given that respect directed to oneself is self-respect, to each of the kinds of personal respect identified in §1 corresponds a kind of self-respect.[123] Nevertheless, my investigation of respect left some issues about self-respect open. In this section I address four such issues. First, I examine

---

[120] There is the further (inessential) reason that the duty of respect, as formulated in footnote 115, does not include the duty of *self*-respect (contrary to the formula of Humanity).

[121] "Ebendarum werden auch die Pflichten gegen den Nebenmenschen aus der ihm gebührenden Achtung nur negativ ausgedrückt, d. i. diese Tugendpflicht wird nur indirekt (durch das Verbot des Widerspiels) ausgedrückt werden" (M6.464.32-465.02/T211; cf. M6.467.28-32/T213).

[122] In this section I consider only (personal) *moral* ASR and RSR.

[123] With the exception of self-comparative AR (footnote 9; cf. §2.3).

the relationships between ASR and self-esteem and between RSR and self-esteem (§3.1). Second, I argue that a conflict between some views about the behavioral content of self-respect can be explained by noticing that "self-respect" is ambiguous between RSR and another kind of self-respect, *standards self-respect* (SSR), which has no straightforward analogue in the case of respect for others (§3.2). Third, I respond to Dillon's (1992c) feminist critique of RSR (§3.3). Fourth, I examine Dillon's (1997) recent proposal that there is yet another kind of self-respect, "basal self-respect" (§3.4).

3.1. Self-respect versus self-esteem

3.1.1. Appraisal self-respect (ASR) versus self-esteem

Like ASR, self-esteem presupposes a positive evaluation of oneself. Unlike ASR, however, self-esteem can depend on "any feature such that one is pleased or downcast by a belief that one has or lacks it"; "for example, one's appearance, temperament, wit, physical capacities, and so forth" (Darwall 1977/1995: 194; cf. Dillon 1995: 30-2).[124] It is thus possible to have high self-esteem and low ASR or vice versa.[125] Given, however, that one's self-esteem is normally greatly affected by one's opinion of one's character, self-esteem and ASR should be positively (though not perfectly) correlated.[126]

---

[124] Even if *moral* ASR is identified with *moral* self-esteem, ASR cannot be identified with self-esteem, because *non*-moral ASR (unlike self-esteem) cannot be grounded on merely natural abilities (footnote 4).

[125] I talk about low ASR rather than appraisal self-*contempt* because it seems that normal people seldom have (appraisal) *self*-contempt. "Low scores on self-esteem scales are typically the result of neutral and intermediate rather than self-derogatory responses to scale items" (Baumeister, Tice, & Hutton 1989: 547).

[126] Like AR, self-esteem can be partial or global (§1.1.1). If I have high ASR but I believe that I'm stupid, or I have low ASR but I believe that I'm physically attractive, then I can have a *fragmented* attitude towards myself (cf. §1.1.1), consisting neither in high nor in low global self-esteem. (On the relation between partial and global self-esteem, see: Marsh 1986, 1993, 1995; Pelham 1995a, 1995b; Pelham & Swann 1989; Tafarodi & Swann 1995.)

3.1.2. Recognition self-respect (RSR) versus self-esteem

Self-esteem can be deficient or excessive; by contrast, one might argue that RSR can be deficient but not excessive. Sachs, for instance, claims that "it is not often thought—indeed it seems a hard saying—that a person has too much [recognition] self-respect" (1981: 347; cf. Dillon 1992b: 128), and that "there may well be no such thing as unwarranted [recognition] self-respect" (1981: 348). Sachs is right if RSR is understood *objectively* (cf. p. 24); i.e., if it is necessary for having RSR that one in fact deserves the treatment which one takes oneself to deserve.[127] RSR can also be understood *subjectively*, however, so that one can have RSR but lack *proper* RSR, and then RSR *can* be excessive. If I mistakenly believe that I deserve to be treated better than others (cf. Boxill 1976/1995: 100), if I protest too loudly against insignificant attacks on my dignity, or if I'm too quick to take offense at relatively harmless teasing, then I may be said to have excessive RSR.[128] In contrast to self-esteem and to ASR, which are clearly subjective, RSR seems to vacillate between a subjective and an objective reading (Massey 1983/1995).[129] It is possible to have high self-esteem and to lack objective RSR or vice versa (Sachs 1981).[130]

---

[127] This definition of objective RSR takes into account the appropriateness of only the cognitive$_3$ component of RSR; one could alternatively define RSR by taking also into account the appropriateness of, e.g., the affective component of RSR (cf. some of the examples that follow in the text).

[128] Dillon (1992b: 128) claims that proper RSR lies between servility and arrogance (cf. Aristotle's doctrine of the mean). Using terminology I introduced earlier (§1.2.4), one could say that objective RSR lies between deferential and condescending subjective RSR. (Note that this is a normative claim.)

[129] Like 'RSR', 'deficient RSR' is ambiguous, and may refer either to absence of objective RSR or to insufficiency of subjective RSR.

[130] Similarly, it is possible to have RSR and to lack RR for others or vice versa (Sachs 1982; cf. Buss 1999b: 538).

3.2. Recognition self-respect (RSR) versus standards self-respect (SSR)

I can believe that I should accede to your request for help either because I believe that you deserve help or because I believe that it would be unworthy of me to refuse. In the former case I focus on *you* as *object* of my behavior, whereas in the latter case I focus on *myself* as *subject* of my behavior. It seems natural to say that in the former case my reason for believing that I should help you is RR for you, whereas in the latter case my reason is self-respect—but what *kind* of self-respect? It cannot be RSR, because RSR (like RR in general) focuses on people as *objects* of behavior, by focusing on the treatment that they deserve. So I claim that it is another kind of self-respect, namely *standards self-respect* (SSR), whose central component is the belief that certain standards of behavior bind oneself, that a certain behavior is worthy or unworthy of oneself.[131]

It might be objected that SSR is just a kind of RSR: if I believe that it would be unworthy of me to refuse your request, then I believe that *I owe it to myself* to accede to your request, and in this sense I consider myself as object of my behavior. But this is not the sense of 'object' that I used in the previous paragraph, where I (implicitly) took an object to be a being towards which the behavior is directed. Moreover, even if talk of subjects and objects of behavior is put aside, the distinction remains between believing that certain standards of behavior bind *P* and believing that *P* deserves to be treated in a certain way. To believe that certain kinds of behavior are unbecoming to an officer is not to believe that the officer deserves a certain treatment. And a monument can deserve to be preserved even though (trivially) no standards of behavior bind it. The distinction can be made even in cases of self-treatment (where one is both the subject and the object of the behavior). I can believe that I should exercise my deliberative capacities either because I believe that my intellect is formidable and thus deserves to be cultivated or be-

---

[131] Note that I understand SSR *subjectively*: in my definition of SSR it doesn't matter whether the standards of behavior *in fact* bind oneself.

cause I believe that, even though my intellect is mediocre, it would be unworthy of me not to make the most of it.

A standard is *self-imposed* (internal) or *externally imposed* according to whether a personal decision (e.g., to abide by the standard) is taken to be or not necessary if the standard is to bind oneself. A self-imposed standard is *personal* or *interpersonal* according to whether it is taken to lack or to have objective authority. Submitting a paper for publication without having first subjected it to the scrutiny of at least three audiences may violate a (self-imposed) standard of mine just because I chose to adopt a corresponding rule; if I view this rule as lacking objective authority, then my standard is personal. What grounds objective authority? Sometimes it's an institution, as when one talks about conduct that would be improper, inappropriate, unbecoming, undignified, unseemly for a military officer or a civic dignitary. Institutional standards are normally self-imposed (and interpersonal, if they are taken to have objective authority), because normally they are taken to bind one only if one has voluntarily adopted them—e.g., by deciding to enter the military.[132] Arguably, however, institutional standards are sometimes externally imposed because they are taken to bind regardless of consent: conscripting people into the military in time of serious war may be legitimate. Externally imposed standards can also be grounded on one's (noninstitutional) status or dignity: it might be claimed that responding to an unprofessionally polemic book review is beneath the dignity of an established but not of a novice writer, or that it's fine for dogs but not for humans to copulate in the street. It's important to notice that the dignity in question, which may be called *subject dignity*, is different from the dignity that grounds WR (*object dignity*). These two kinds of dignity are easily confused, maybe because of the

---

[132] Institutional standards normally *bind* one only if one has voluntarily adopted them because they normally *apply* to one only if one has voluntarily adopted them. (A norm—or standard—*applies* to one exactly if one falls under the scope of the norm; a norm *binds* one exactly if it applies to one and it has objective authority.)

common (normative) belief that humans have both, but one can imagine examples in which they come apart. A hierarchical society is possible in which members of the lowest caste are believed to have no object dignity (there is no treatment that they are believed to deserve, even from themselves) but to have subject dignity (some conduct is considered unworthy of them—indeed of every human being).

The importance of standards for self-respect has been recognized by several philosophers (Dillon 1992b; Hill 1985/1995b; G. Taylor 1985; Telfer 1968/1995).[133] Attending to the distinction between RSR and SSR illuminates a conflict about the behavioral content of self-respect. Does self-respect enjoin *protest* when one's rights are violated by an insensitive oppressor? (Boxill 1976/1995.) Focusing on RSR suggests a positive answer: since I deserve a better treatment, why shouldn't I claim that I deserve it? Focusing on SSR, however, suggests a negative answer: if I know that protesting would be to no avail, then it may be undignified for me to protest.[134]

3.3. Reply to Dillon's feminist critique of recognition self-respect

Dillon (1992c; cf. 1992a) has criticized RSR on feminist grounds. Dillon claims that to "have recognition self-respect is to take appropriate account of the fact that one is a person, to properly appreciate one's worth as a person" (1992c: 56), and then complains that this "taking-account-of-and-appreciating attitude is a dispassionate, overly

---

[133] Dillon's (1992b: 133-4) "interpersonal recognition self-respect", "agentic recognition self-respect", and "personal recognition self-respect" apparently correspond, respectively, to RSR, SSR with externally imposed standards, and SSR with personal self-imposed standards. (Given that SSR has no straightforward analogue in the case of respect for others, I prefer to say that RSR and SSR are two kinds of self-respect, rather than saying—as Dillon does—that they are two kinds of a broadly construed "recognition" self-respect.) Telfer's (1968/1995) "conative self-respect" apparently corresponds to SSR, and so does the kind of self-respect that Hill (1985/1995b) identifies.

[134] Protest is *outward* opposition (to the violation of one's rights); thus refraining from protesting is compatible with *inward* opposition and with RSR (cf. Sachs 1982: 124). Boxill (1976/1995) argues that I should nevertheless protest in order to reassure myself that I have self-respect. Rawls (1971: 440), on the other hand, claims that self-respect presupposes self-confidence.

intellectualized, arm's-length response that does not engage us emotionally" (1992c: 58). But Dillon's complaint is directed towards an impoverished, purely cognitive concept of RSR. If a disposition to feel RSR is necessary for having RSR (cf. p. 5), then Dillon's complaint loses its sting.

Dillon's main target is what she calls the "abstractive conception" of RSR: "in viewing us as worthy of respect", RSR "abstracts from all particularities, regarding the details of ourselves as irrelevant to our intrinsic moral worth" (1992c: 56; cf. 1992a: 116; Noggle 1999: 454). "However, it is difficult to understand how regarding oneself in generic terms could constitute *self*-respect". Moreover, RSR "so understood is compatible with and perhaps even encourages self-alienation, for it allows that I can respect myself without paying attention to who I am, without taking *me* seriously" (1992c: 57). I reply that "regarding the details of ourselves as irrelevant to our intrinsic moral worth" need not prevent us from responding to *our* worth when we have RSR. Take an analogy. Suppose that, while browsing in a library, I come upon a rare book and I have a spontaneous tendency to handle it with care. My belief that all rare books deserve to be handled with (equal) care need not prevent my tendency to handle *this* book with care from being a response to what I take to be *this* book's value (hence a response to *this* book). Similarly, I have WR (or RR) for someone (including myself) only if the affective and conative components of my attitude are responses to what I take to be *this* person's worth (cf. footnote 19), rather than, e.g., to my belief that all persons (equally) deserve a certain treatment. Dillon's claim that "I can respect myself without paying attention to who I am" is in a sense true[135] but is in the present context misleading: I *am* paying attention to

---

[135] Namely, in the sense that I am abstracting from all particularities that distinguish me from other persons.

the (putative) fact that I am a being with worth, so that I *am* taking myself seriously, and the charge of self-alienation is defused (cf. Velleman 1999: 367-8).[136]

Some of Dillon's criticisms are directed not towards RSR *per se*, but rather towards RSR in conjunction with a conception of persons "as essentially rights-bearers". If "to respect a person's rights is to keep one's distance from her", then RSR may "make it difficult to envision ourselves as being-in-relation with others and to value ourselves as connected with others"[137] (1992c: 57). "[T]o the extent that the self-conceptions of many women *do* include regarding themselves as selves-in-relation, ... such an understanding of self-respect does not take seriously the actual concerns of many women" (1992c: 58). I have two replies. First, insofar as RSR (like RR) focuses on a fundamental feature that we are supposed to share with everyone, namely on the (putative) fact that we all have equal intrinsic worth, RSR does promote a conception of ourselves as selves-in-relation. Second, as I explained in §1.2.1, positive treatment need not be understood as noninterference, but can be understood as promoting the good of others, in which case Dillon's criticism is rebutted.[138]

---

[136] A different response to Dillon's charge of self-alienation was suggested to me by Stephen Darwall, who claimed that I have RR for someone (including myself) only if my belief that this person deserves a certain treatment is a response to what I take to be *this* person's worth, rather than the outcome of an inference from the generalization that every person deserves a certain treatment. But Darwall's response leaves open the possibility that the affective and cognitive components of my attitude are responses to my belief that this person deserves a certain treatment, rather than to what I take to be this person's worth; thus Darwall's response is open to "one thought too many"-style objections (cf. footnote 19). Moreover, I disagree with Darwall's claim: I think I can have RR for you even if my belief that you deserve a certain treatment is the outcome of a laborious inferential process. Returning to my analogy, if I realize only after careful examination that the book I came across is rare, my subsequent tendency to handle this book with care can still be a reaction to what I take to be *this* book's value, rather than to my belief that this book deserves to be handled with care.

[137] Actually, it seems more appropriate to regard this criticism as directed towards RR for others than towards recognition *self*-respect.

[138] Similar replies can be given to two other criticisms made by Dillon, namely (a) that "the oppositional, combative conception of the relations among persons that is encouraged by construing respect as a defense of rights is not conducive to the formation of mutually supportive and integrative relationships among persons" (1992c: 58-9), and (b) that "the focus on defensiveness promotes a preoccupation with victimization—valuing oneself as victim—that is far from empowering" (1992c: 59).

3.4. Dillon's "basal self-respect"

Dillon asks us to consider the cases of three "real women" (1997: 235), whom she calls "Anne", "Beth", and "Carissa". "Anne is a successful professional" who "*knows* that she deserves to take pride in her accomplishments" and "*believes* she *is* respect-worthy" but "feels totally inadequate and undeserving". Beth "regards pride as a fitting response to her embodied being" but still "feels dirty when she menstruates and can't look at her naked body without disgust". Carissa "routinely feels resentment about things she cannot reasonably regard as disrespectful" (e.g., "she resents telephone solicitors"), although "she knows her resentment is ungrounded" (1997: 232-3). Dillon suggests that "the distortions of recognition and evaluative self-respect played out in Anne's, Beth's, and Carissa's anomalous emotions arise from damaged basal self-respect" (1997: 241). Dillon takes basal self-respect to be "a more fundamental orientation towards the self that underlies recognition and evaluative self-respect, a prereflective, unarticulated, emotionally laden presuppositional interpretive framework, an implicit 'seeing oneself as' or 'taking oneself to be' that structures our explicit experiences of self and worth" (1997: 241).

Contrary to the kinds of (self-)respect that I have examined so far, whose existence is supposed to be made obvious by introspection, basal self-respect is an *inferred*, theoretical construct, whose existence is *postulated* in order to account for certain observable phenomena. Let us be clear on this point: postulating the existence of basal self-respect is needed, if at all, only in order to *explain* the psychological processes in the cases of Anne, Beth, and Carissa. In order to *describe* these cases it's unnecessary to use basal self-respect: it suffices to use the distinctions between real and apparent beliefs and between attitudes and feelings. Anne, for instance, has the *apparent* belief that she is worthless and the *real* belief that she is valuable; she *feels* but she does not *have* (non-moral) appraisal self-contempt (§1.1.2); she does not feel or have but she believes that she *should* feel and have ASR. So it seems that, if we are to accept Dillon's hypothesis

that basal self-respect exists, we must be shown not (only) that this hypothesis provides a *possible* explanation of the phenomena that Dillon describes, but that it provides the *best* possible explanation; and this Dillon does not even attempt to show.

There is also reason for being skeptical of Dillon's hypothesis. Dillon apparently takes basal self-respect to be a unitary, global preconscious "attitude" towards oneself, whereas it seems possible to have fragmented preconscious attitudes towards oneself. Suppose, for instance, that Beth, besides feeling disgust for her body, feels proud of her intellect, even though she believes that her intellect is mediocre; does she have then damaged basal self-respect or not?

I conclude that Dillon's hypothesis that basal self-respect exists stands in need of further support.

CHAPTER II
AN EPISTEMIC CRITIQUE OF GLOBAL APPRAISAL RESPECT

In Chapter I we saw that both Kant and contemporary philosophers distinguish between appraisal and recognition respect. In this chapter I present an epistemic critique of global (personal moral) appraisal respect, an attitude based on a positive appraisal of a person as having a good character. In §1 I give an outline of my argument. In §2 I use empirical evidence from social psychology (including Milgram's obedience experiments) to argue that most people are *fragmented*: they (would) behave deplorably in many and admirably in many other actual or counterfactual situations. In §3 I argue that, according to certain plausible conceptions of character evaluations, being fragmented entails being *indeterminate*: neither good nor bad nor intermediate. It follows that *most* people are indeterminate. In §4 I use empirical evidence from personality psychology to argue that our information about specific people almost never distinguishes those who are indeterminate from those who are not. It follows that evaluations of people as good, bad, or intermediate are almost always epistemically unwarranted. I conclude that appraisal respect and contempt are epistemically criticizable as based on unwarranted beliefs. In §5 I address objections to my argument. In §6 I relate my argument to recent work by John Doris and Gilbert Harman. In §7 I conclude with a proposal to replace global character evaluations with *local* evaluations of people in light of their behavior in relatively narrow ranges of actual and counterfactual situations.[1]

---

[1] Material from this chapter was presented at the seventy-third annual meeting of the Pacific Division

1. Outline of the argument

My main thesis in this chapter (call it the *epistemic thesis*) is that character evaluations are almost always epistemically unwarranted, *unjustified*.[2] This is not to deny that sometimes they are *accurate* (some unjustified beliefs are true), so the thesis is not that no good or bad people exist. The epistemic thesis is in a way analogous to the claim that one is almost never justified in believing that one's lottery ticket will win—even if the ticket does happen to win. The thesis is not about *expressing* evaluations; it is about *making* (i.e., forming or holding) evaluations, even if they are never expressed. The thesis is about evaluations of *character*, not of character *traits* like honesty or courage. Nor is the thesis about evaluations of *actions* as good or bad; in fact I will *use* such action evaluations to support the epistemic thesis. If the epistemic thesis is true, then feelings and attitudes like those of global appraisal respect (i.e., esteem) and contempt, which are based on character evaluations, are epistemically criticizable: similarly to the way in which one's fear of flying is epistemically criticizable if it is based on an unjustified be-

---

(Berkeley, April 1999; commentator: Lyle Zynda) and at the ninety-seventh annual meeting of the Central Division (New Orleans, May 1999; commentator: Jami Anderson) of the American Philosophical Association, as well as at the University of Michigan (Decision Consortium Conference, May 1999, Department of Philosophy, September 1999, and Judgment and Decision Making seminar, November 1999), Carnegie Mellon University (January 2000), New York University (January 2000), California Institute of Technology (May 2000), Iowa State University (January 2001), University of St Andrews (February 2001), and Jesus College, Oxford (March 2001). In addition to the debts I mentioned in the Acknowledgments, I am indebted to the following people for interesting questions: Steve Awodey, Gordon Belot, Karen Bennett, Michael Bishop, Daniel Bonevac, John Broome, Philip Buckley, Travis Butler, Colin Camerer, William Copeland, Fiona Cowie, Ann Cudd, Garrett Cullity, Justin D'Arms, Steven Daskal, Kevin de Laplante, Janine Diller, Craig Duncan, Carla Fehr, Hartry Field, Samuel Floyd, Tamar Gendler, Gyl Gentzler, Heimir Giersson, Clark Glymour, Steven Gross, Bart Gruzalski, Alan Hájek, John Haldane, Gary Hatfield, Rodney Hayward, Thomas Hill, Christopher Hitchcock, Bob Hollinger, Margaret Holmgren, Cory Juhl, Frances Kamm, Kevin Kelly, Amy Kind, Celery Kovinsky, Rahul Kumar, Joe Kupfer, Alison Laywine, Fraser MacBride, Kathleen McShane, Gerald Massey, George Mavrodes, Dominic Murphy, Thomas Nagel, Stephen Phillips, Gerald Postema, Steven Quartz, George Rainbolt, John Richardson, Michael Ridge, Samuel Ruhmkorff, Stephen Schiffer, Tommie Shelby, John Skorupski, Tony Smith, Samuel Sommers, Peter Spirtes, Robert Stalnaker, Gabrielle Starr, Peter Unger, Peter Vanderschraaf, Jonathan Vogel, Nicholas White, James Woodward, and Frank Yates.

[2] I use "character evaluations" to refer interchangeably to evaluations *of character* and to evaluations *of people* in terms of their character (as good, bad, or intermediate). I am talking about *moral* character.

lief that flying is much more dangerous than driving, one's esteem for a person is epistemically criticizable if it is based on an unwarranted evaluation of the person as good.

The epistemic *thesis* should be now relatively clear; but what is my *argument* for this surprising thesis? Think again about the lottery analogy. Most tickets will not win, so my *prior* probability that my ticket will not win should be high. My information about my ticket (e.g., information about its color or number) should almost never change this probability (except, e.g., if I know that the lottery is rigged), so my *posterior* probability that my ticket will not win should again be high. It follows that I'm almost never justified in believing that my ticket will win. My argument for the epistemic thesis has a similar structure. First, I argue that most people are what I call *indeterminate*—neither good nor bad nor intermediate—so that our *prior* probability that any given person is indeterminate should be high. Second, I argue that our information about specific people (e.g., information about their behavior in various situations) should almost never appreciably lower this probability, so that our *posterior* probability that any given person is indeterminate should again be high. It follows that we are almost never justified in evaluating a person as good (or bad, or intermediate). But now I have exchanged one surprising thesis for two. Why do I claim that most people are indeterminate? And why do I claim that our information about specific people is almost never sufficiently informative? I will briefly address these two questions in turn; in later sections I provide details and I discuss numerous objections.

First, why do I claim that most people are indeterminate? Because I claim that (1) most people are what I call *fragmented*—they (would) behave deplorably in many and admirably in many other actual or counterfactual situations—and that (2) being fragmented entails being indeterminate. Here are some preliminary details. (1) My claim that most people are fragmented is based primarily on an extensive literature in social psychology which suggests that most people can be induced to behave deplorably (or admirably) in various experimental contexts. One striking example is Milgram's experiment,

which has been replicated all over the world, and in which most participants were induced to administer increasingly powerful (in fact fictitious) electric shocks to a writhing and screaming middle-aged man (in fact a confederate of the experimenter; details and further examples in §2). (2) My claim that being fragmented entails being indeterminate is based on the claim that character evaluations presuppose at least moderate consistency in people's behavior; good people, for example, are expected to resist all but a few temptations to behave deplorably. *Perfect* consistency we don't expect, but *acute* inconsistency baffles us: a serial rapist who regularly saves lives as a volunteer at a suicide prevention center eludes character classification. Such a person is not *between* good and bad. He is in a sense *both* good and bad, and so is simpliciter neither: he is indeterminate. Such inconsistency is of course exceptional, but in §3 I argue that less extreme and even *counterfactual* inconsistency suffices for indeterminacy. In short, character evaluations presuppose at least moderate behavioral consistency, so that fragmentation (i.e., acute inconsistency) entails indeterminacy; but fragmentation, surprisingly, is the statistical norm; thus indeterminacy, surprisingly, is also the statistical norm.

Second, why do I claim that our information about specific people should almost never appreciably lower our probability that they are indeterminate? Because I claim that (1) our information could do so only if what I call the *Independence Condition* were violated, and that (2) the Independence Condition holds. The Independence Condition says (roughly) that people's behaviors in various situations should be considered approximately independent of each other; e.g., learning that a person behaved deplorably in situation 1 should not significantly affect our confidence that the person would behave deplorably in situation 2. In §4 I support this condition by appealing to an extensive literature in personality psychology which suggests that from information on how people behave in certain situations one cannot confidently predict how they (would) behave in other, dissimilar situations. I am willing to grant that in truly exceptional cases our in-

formation may warrant such predictions; this is why the epistemic thesis says that character evaluations are *almost* always epistemically unwarranted.

Let me summarize and state more formally the argument for the epistemic thesis that I previewed in the last three paragraphs (and that I elaborate in the next three sections). $P(Ip)$ is the (prior) probability that person p is indeterminate, and $P(Ip|Ep)$ is the conditional (posterior) probability that p is indeterminate given that our information (evidence) about p is E.

(Q1) Most people are fragmented.

(Q2) Fragmentation entails indeterminacy.

Thus (from Q1 and Q2): (L0) Most people are indeterminate.

Thus (from L0): (L1) For any p, $P(Ip)$ should be high.

(P1) For any p, if the Independence Condition holds, then $P(Ip|Ep)$ should *not* be appreciably lower than $P(Ip)$.

(P2) For almost any p, the Independence Condition holds.

Thus (from P1 and P2): (L2) For almost any p, $P(Ip|Ep)$ should *not* be appreciably lower than $P(Ip)$.

Thus (from L1 and L2): (C1) For almost any p, $P(Ip|Ep)$ should be high.

(Thus $P(Gp|Ep)$ and $P(Bp|Ep)$, the posterior probabilities that p is good or bad, should almost always be low.)[3]

In §2-§5 I defend the epistemic thesis: in §2 and §3 I deal with the *prior* probability of indeterminacy by defending respectively Q1 and Q2, in §4 I deal with the *posterior* probability of indeterminacy, and in §5 I address objections to the epistemic thesis.

---

[3] The above argument for the epistemic thesis is suited for expository purposes but differs slightly from the argument I will actually defend; see §4.1 for details. Note that $P(\neg Gp|Ep)$ should be high if $P(Gp|Ep)$ should be low, so on my view evaluations of people as *not* good (similarly, as *not* bad) are almost always epistemically *warranted.*

2. Most people are fragmented (Q1)

2.1. The concept of fragmentation and the argument for Q1

I call a person *fragmented* exactly if the person does or would behave deplorably in an open list of actual or counterfactual situations and admirably in another such open list. I understand an *open list* of situations as comprising an indefinitely large (though not necessarily infinite) number of multifarious situations. I call an action (token) *deplorable* when it is seriously blameworthy[4] and *admirable* when it is highly praiseworthy. An action is *blameworthy* or *praiseworthy* exactly if its performance makes the agent deserve blame or praise respectively; alternatively (and, I suggest, equivalently), an action is blameworthy exactly if it is wrong (in the sense of violating one's duty) and lacks an adequate excuse,[5] and is praiseworthy exactly if it is supererogatory (in the sense of exceeding one's duty) and lacks a "defeater" (e.g., an ulterior motive).[6] It follows that whether an action is deplorable or admirable may depend not only on its consequences, but also on the agent's motives, intentions, beliefs, and so on. I don't need to provide a

---

[4] 'Deplorable' can mean (a) 'lamentable', 'regrettable' or (b) 'shameful', 'disgraceful' (Urdang 1992: 95); I adopt the latter use.

[5] I am not taking a stand on whether wrongness is a more basic notion than blameworthiness or vice versa: my suggested equivalence between being blameworthy and being wrong in the absence of an adequate excuse (schematically: $B \leftrightarrow W \wedge \neg E$) is compatible with an equivalence (inspired by Gibbard 1990b: 45) between being wrong and being blameworthy unless adequately excused (schematically: $W \leftrightarrow (\neg E \rightarrow B)$). In fact, it can be seen that my suggested equivalence amounts to the conjunction of (1) $W \rightarrow B \vee E$, (2) $B \rightarrow W$, and (3a) $E \rightarrow \neg B$, whereas the other suggested equivalence amounts to the conjunction of (1), (2), and (3b) $E \rightarrow W$; both (3a) and (3b) seem true. One might object to (3a) that "few excuses get us out of it *completely*" (J. L. Austin 1956/1979: 177); for example, a rape can be blameworthy even if it is excused by the fact that the agent's upbringing made him into a serial rapist. In reply I may grant that few excuses are *adequate*: the agent's upbringing may provide a *mitigation* (cf. Sabini & Silver 1982: 151 n. 10), a partial (inadequate) excuse.

[6] Contrary to Feinberg (1961/1970a: 12), and probably deviating from common usage, I understand supererogation as not including praiseworthiness: if you risked your life to save your uncle from drowning only to make him include you in his will, I would call your action supererogatory but not praiseworthy. (Even those who do not share this terminological preference may need a term for actions which exceed one's duty regardless of (e.g.) how they are motivated—which is how I am using 'supererogatory'.) Since I understand praiseworthiness as including supererogation, I take it that people don't *deserve* praise just for doing their duty, even if in some cases it *makes sense* to praise them. (For example, it may make sense for me to praise you for doing your duty in a situation in which doing one's duty is very difficult and rare.)

general account of this dependence: for my purposes it will suffice to support my specific claims that certain actions are deplorable or admirable. Fragmentation is a property that a person can have during some time periods and lack during others: by definition, only *current* (actual or counterfactual) behavior is relevant to whether a person is currently fragmented. My definition of fragmentation makes no presuppositions about *why* the agent behaves sometimes deplorably and other times admirably; in particular, the definition does not presuppose that the agent has a "modular mind" (Fodor 1983) or a "fragmented psyche" consisting of good and evil parts interlocked in a Manichaean struggle.

Having thus clarified the concept of fragmentation, I give now my argument for the claim that (Q1) most people are fragmented.

(Q3) There are many situations in each of which most people (would) behave deplorably.

(Q4) There are many situations in each of which most people (would) behave admirably.

Thus: (Q1) Most people (would) behave deplorably in many (i.e., in an open list of actual or counterfactual) situations and admirably in many other situations.

The validity of this argument is not obvious; I examine it in §2.4 after I defend successively Q3 (§2.2) and Q4 (§2.3).

2.2. Situations in which most people would behave deplorably (Q3)

I will defend Q3 by examining three (kinds of) experiments from social psychology: (1) Milgram's experiments (in which most participants administered powerful—in fact fictitious—electric shocks to a screaming confederate), (2) Zimbardo's experiment (in which most "guards" in a simulated prison maltreated the "prisoners"), and (3) some bystander intervention experiments (in which most participants failed to help an apparent seizure victim). For each experiment I will argue that (I) most *participants* behaved de-

plorably and that (II) most *people* would behave as those participants who behaved deplorably did. Clearly, the existence of *three* situations in which most people would behave deplorably does not (deductively) entail the existence of an *open list* of such situations. Nevertheless, I think that my argument has high inductive strength, because it suggests that with more ingenuity and less ethical scrupulosity (combined with decreased restrictions from Institutional Review Boards for the Protection of Human Subjects) social psychologists could extend indefinitely the list of such situations.

2.2.1. The obedience experiments

Milgram's (1974) nineteen *obedience experiments*,[7] conducted between 1960 and 1963, are all variations on the same theme; I will describe one basic variant, which I will be referring to as 'Milgram's experiment'. Upon arrival at the Yale University laboratory, the participant (who has come in response to a letter or newspaper advertisement offering $4.50 for a one-hour participation in a "study of memory") meets another putative participant (in fact a confederate of the experimenter). The experimenter states that the purpose of the study is to investigate the effects of punishment (specifically, by administering electric shocks) on learning. After a rigged draw which assigns to the participant the role of "teacher" and to the confederate the role of "learner", the confederate is strapped into an "electric chair" apparatus. He claims to suffer from a slight heart condition, nothing serious, and asks whether the shocks are dangerous; the experimenter replies that the shocks can be extremely painful but cause no permanent tissue damage. The participant then goes to an adjacent room and sits in front of a "shock generator" having 30 switches labeled from 15 to 450 volts (in 15-volt increments); groups of

---

[7] Milgram's *studies* involve no random assignment of participants to conditions and are thus not *experiments* in the strict sense of the term 'experiment' in social science research (e.g., Aronson, Ellsworth, Carlsmith, & Gonzalez 1990: 17). Nevertheless, it is common to call informally Milgram's studies 'experiments' (e.g., Aronson et al. 1990: 70, 89).

switches bear verbal designations which include "Extreme Intensity Shock" and "Danger: Severe Shock". The participant is given a sample shock of 45 volts to strengthen his belief in the authenticity of the generator. The experimenter instructs the participant to administer a shock to the learner each time the learner gives a wrong answer in the learning task, starting from 15 volts and *moving one level higher* on the generator with each wrong answer. The learner receives in fact no shocks but goes through a standardized sequence of protests (played back from tape): he claims that his heart is bothering him, demands insistently to be released, refuses to answer further questions, and emits progressively longer and stronger agonized screams. After 330 volts, he is no longer heard from. The experimenter remains impassive throughout; he instructs the participant to treat a lack of answer as a wrong answer, refuses to humor any request to check on the learner's condition in the next room, and reiterates that the shocks cause no permanent tissue damage. If the participant refuses to continue, the experimenter goes through a sequence of four "prods", the last two being "it is absolutely essential that you continue" and "you have no other choice, you *must* go on". If the participant disobeys after the fourth prod, the experiment stops; otherwise, the experiment continues until the participant depresses the 450-volt switch three times.

(I) Call those participants who went *at least* up to 225 volts (i.e., halfway through the "shock generator") *obedient* and those who went all the way up to 450 volts *fully obedient.* Call those participants who were certain or at least highly confident that the learner was getting painful shocks *nonsuspicious* and the remaining participants *suspicious*. My claim that most participants behaved deplorably is based on two premises: (i) most participants were nonsuspiciously obedient; (ii) nonsuspicious obedience was deplorable. The first premise is based on the facts that 80% of the 40 participants were obedient (65% were fully obedient) and that according to a follow-up questionnaire about 80% of the participants were nonsuspicious (so that about 64%—i.e., 80% of

80%—of the participants were nonsuspiciously obedient).[8] The second premise is based on the claims that nonsuspicious obedience was not adequately excused (see below) and that it was seriously wrong: it violated the duty to avoid acting so as to inflict severe pain on an innocent and nonconsenting person.[9] Both premises are subject to powerful objections.

Objecting to the first premise, one might claim that the incongruity between the experimenter's imperturbability and the learner's apparently extreme suffering must have made most participants seriously doubt that the learner was getting shocks (Orne & Holland 1968: 287).[10] I have two replies. First, participants who relied on the experimenter's reassurance that the shocks were not dangerous may have interpreted the experimenter's imperturbability as due to a blasé attitude (much like some dentists can be blasé about

---

[8] The 64% figure is approximate for at least three reasons. First, Milgram (1974: 172) gives no data for what I call 'obedient' participants: he gives data for all participants (80.1% nonsuspicious), fully obedient participants (73.8% nonsuspicious), and remaining participants (85.1% nonsuspicious). Second, Milgram's data lump together participants from many experiments. Third, maybe some suspicious participants lied in the follow-up questionnaire in order to please the experimenter or in order to avoid spoiling an expensive and time-consuming scientific experiment (Orne 1962: 780; Orne & Holland 1968: 285; Patten 1977b: 432). To support this possibility, Orne and Holland refer to a study in which "three quarters of the [participants] ... indicated that they did not really believe the deception when carefully questioned after the experiment" (1968: 290). On the other hand, Milgram refers to a study in which, "[o]n the basis of highly stringent criteria of full acceptance, ... 60 percent of the subjects fully accepted the authenticity of the experiment" (1974: 173). I find dubious the above reasoning according to which a suspicious participant would lie. A participant smart enough to see through the attempted deception will presumably also be smart enough to realize that an experimenter who asks "did you believe that the learner was getting shocks?" both *hopes* to hear a positive answer and *wants* to hear a truthful even if negative answer; therefore, a wish to please the experimenter does not ensure that such a participant will answer positively. Moreover, a wish to avoid spoiling the experiment will arguably make such a participant answer negatively. Note also that maybe some *non*suspicious obedient participants lied in the questionnaire in order to make their obedience look excusable and thus appear to the experimenter in a positive light (cf. Berkowitz & Donnerstein 1982: 250).

[9] Patten objects that "a nurse or dental assistant ... might inflict severe pain upon command of a doctor or dentist where we would not be tempted to say there was anything unethical about the action" (1977a: 352; cf. Martin, Lobb, Chapman, & Spillane 1976: 354). But in such cases there is at least implicit consent; moreover, in such cases the inflicted pain is for the patient's own good, not for the good of science (Pigden & Gillet 1996: 248).

[10] Cf. Mixon 1972: 159. However, in a variant of Milgram's experiment in which the experimenter was not impassive, 52 of the 57 participants were fully obedient (Ring, Wallston, & Corey 1970).

patients' screams). Second, if suspicions among participants were widespread, then how come most participants protested repeatedly or "were observed to sweat, tremble, stutter, bite their lips, groan, and dig their fingernails into their flesh"?[11] Orne and Holland (1968: 287) respond with an analogy: in a stage magician's trick in which a volunteer from the audience is strapped into a guillotine and another volunteer is requested to trip the release lever, the latter volunteer is likely to feel nervous despite knowing that it's only a trick. I reply that this analogy fails on two counts. First, the volunteer is unlikely to protest or disobey the magician's request, whereas most participants protested and many eventually disobeyed the experimenter's requests. Second, the volunteer will probably feel only mild nervousness, whereas many participants displayed severe nervousness.[12] Given these differences between the nervousness of the participants and that of the volunteer, it is plausible to explain the former—even if not the latter—by appealing to a belief about pain or harm.

Objecting to the second premise, one might claim that nonsuspicious obedience was not deplorable if it was based on *justified trust* in the experimenter (Harré 1979: 105; Mixon 1989: 29, 41). But in what exactly would such trust consist? Not in the belief that the learner was not getting shocks, since we are talking about *non*suspicious obedience.[13] Maybe the trust consisted in the belief that the experimenter had some

---

[11] Milgram 1963: 375; cf. 1965b: 66, 1972: 139-40; Modigliani & Rochat 1995: 117; Rochat, Maggioni, & Modigliani 2000: 168-70. These reactions don't exonerate the participants: one can behave deplorably even if one behaves unwillingly. These reactions also provide a reply to Mixon's (1989: 37-8) claim that the learner's protests and screams, as featured in Milgram's (1965a) film, are unconvincing.

[12] One might object that in a "role-playing" variant of Milgram's experiment (in which the participants were told that the learner would not get shocks but were asked to behave as if they had not been told) most (though not all) participants displayed severe nervousness (Mixon 1972: 150). I reply that in another role-playing variant this result was not replicated: most participants displayed only mild nervousness (O'Leary, Willis, & Tomich 1970: 91).

[13] One might object that I defined nonsuspicious participants as those who were certain *or highly confident* that the learner was getting painful shocks, so that *some* nonsuspicious participants doubted that shocks were actually administered (cf. Patten 1977b: 431). I reply that "it would surely be right not to operate the shock machine at all rather than to take even a slight risk of inflicting pain on a person" (Ingram 1979:

(perhaps unfathomable) *scientifically* valid reason for conducting the experiment.[14] I reply that such trust, even if epistemically warranted, would not morally justify nonsuspicious obedience because it would not guarantee that the experimenter also had a *morally* valid reason for asking the participants to inflict severe pain on the learner: even if experiments are normally scientifically justified, they need not always be morally justified. (Witness Sheridan and King's variant of Milgram's experiment in which 20 out of 26 participants were fully obedient in administering *real* shocks to a "running, howling, and yelping" "cute, fluffy puppy" (1972: 165), or Landis's experiment in which 15 out of 21 participants, "after more or less urging" (1924: 459), complied with the experimenter's request to behead a *live* white rat with a butcher's knife![15]) Maybe, however, the trust included in addition the belief that the experimenter had some such morally valid reason.[16] I reply that such trust would at most *explain* but would again not morally *justify*

531; cf. Coutts 1977: 520; Darley 1995: 133)—let alone a *considerable* risk (in case the doubt was only slight; cf. Pigden & Gillet 1996: 237).

[14] There is indeed evidence that such trust is prevalent in experimental situations. Orne reports that, when asked to perform serial additions on sheets filled with rows of random digits and to tear up each completed sheet into a minimum of 32 pieces before going on to the next sheet, participants continued for several hours until the *experimenter* gave up! In postexperimental interviews, participants would "invariably attribute considerable meaning to their performance, viewing it as an endurance test or the like" (1962: 777; contrast Frank 1944: 24).

[15] "All subjects argued about the following of directions, most of them doubting the experimenter's sincerity in the demand to actually kill the rat. There was a great deal of vacilliation [sic], many false starts and then a final hurried reaction which because of the effort and attempt to hurry usually resulted in a rather awkward and prolonged job of decapitation" (Landis 1924: 486).

[16] There is again evidence that such trust is prevalent in experimental situations. In an experiment by Orne and Evans (1965), participants who were simulating hypnosis watched a coin dissolve in fuming nitric acid for about a minute and then, at an experimenter's request, picked the coin out of the acid with their bare hands! The participants afterwards said that they assumed it was safe to do so (it was!) because they took the experimenters to be competent and responsible. On the other hand, the existence of such trust may not fully explain obedience in Milgram's experiment. First, because most participants protested. Second, because most participants were obedient even when such trust may have been undermined. For example, 26 out of 40 participants were obedient (19 were fully obedient) in a variant of Milgram's experiment which was conducted not under the auspices of Yale University, but rather in a "marginally respectable" laboratory "in a somewhat rundown commercial building" in Bridgeport, Connecticut, supposedly by "a private firm conducting research for industry" (Milgram 1974: 68-9). Cf. also Mantell's (1971: 104-6) "modeling delegitimization" variant, in which the participants watched a confederate disobey after the experimenter

nonsuspicious obedience because it would be epistemically unwarranted: the participants' belief that the learner was in agony should have made them question the experimenter's moral (as opposed to scientific) competence (Pigden & Gillet 1996: 248; cf. Patten 1977a: 363). One might rejoin that the participants relied on the experimenter's reassurance that the shocks, although painful, were not dangerous (Mixon 1989: 32). I reply that the perceived pain itself should have made the experiment look morally unacceptable (and did make nonsuspicious obedience deplorable) even in the absence of any perceived danger (Ingram 1979: 532; Pigden & Gillet 1996: 247; cf. Darley 1995: 128, 134). (Moreover, there was arguably reason to perceive some danger of heart trouble; cf. Hamilton 1992.)

One might also object to the second premise by claiming that an action which most people perform cannot be deplorable.[17] In reply I grant that *some* actions which *almost everyone* performs are excusable: maybe it's not deplorable to divulge state secrets when tortured in a way that makes almost everyone succumb. But nonsuspicious obedience in Milgram's experiment was nowhere near universal: a substantial minority did disobey (cf. Miller 1995: 47). Indeed, a nonsuspicious obedience rate of about 64% seems tailor-made for my purposes: it corresponds to a majority, but not to a majority so overwhelming as to make plausible the claim that nonsuspicious obedience was excusable.[18] Moreover, some actions are deplorable although almost everyone performs them:

"admitted" that he was a student and not a member of the Max Planck Institute and that the experiment was not being supervised, but 52% of the 25 participants were still fully obedient.

[17] A related claim is that most people would consider nonsuspicious obedience to be nondeplorable if they learned that most participants were nonsuspiciously obedient. Even if true, this claim would not *contradict* the second premise (which says that nonsuspicious obedience *was* deplorable, not that it would be *considered* deplorable by most people); this claim might be *evidence* against the second premise, but this evidence would be outweighed by my argument in favor of the second premise.

[18] One might object that almost everyone (97.5% of the participants) went at least up to 150 volts (cf. Harman 1999: 322). I have two replies. First, only about 78% (i.e., 80% of 97.5%) of the participants did so *non*suspiciously. Second, given that the learner did not withdraw his consent until 150 volts, arguably going nonsuspiciously up to 150 volts was not deplorable even if going nonsuspiciously up to 225 volts was.

consider the inactivity of 38 witnesses to the Kitty Genovese murder (A. M. Rosenthal 1964), or the "selections" performed by German doctors in Nazi concentration camps. (The selections consisted in deciding who would be allowed to live for a while and who would be immediately sent to the gas chambers; see Lifton 1986: chap. 8-11.) It seems thus that the excusability of an action is not guaranteed by near-universal performance of the action but depends rather on features of the situation. I don't need to provide a general account of this dependence: for my purposes it suffices to point out that Milgram's experimental situation did not make nonsuspicious obedience excusable because no dire consequences threatened disobedient participants.

In response one might argue that the experimental situation had several mitigating features. (i) The participants came unprepared for the possibility that they would face a morally problematic situation, and the experiment was so fast-paced that they had little time to reflect: they "acted without choosing" (Bok 1996). I reply that an action can be deplorable even if performed on the spur of the moment. (ii) The stepwise nature of the experiment made it hard to disobey at any particular point of the "shock generator" given that one had obeyed at the immediately preceding point (Gilbert 1981; cf. Meeus & Raaijmakers 1995: 159; Sabini & Silver 1982: 70; Shanab & O'Neill 1979: 242). I reply that there was a natural disobedience point: 150 volts, when the learner first withdrew his consent to continue (cf. L. D. Ross 1988: 103). (iii) The participants may have believed that it was illegitimate of the learner to withdraw his consent (Mantell & Panzarella 1976: 243). I reply that such a belief was at most a very partial explanation of the participants' obedience, given that 70% of participants were obedient in a variant of Milgram's experiment in which the learner explicitly stated that he agreed to participate "only on the condition that the experiment be halted on his demand" (Milgram 1974: 64). (iv) The participants had freely consented to participate and thus felt an obligation to comply with the experimenter's requests (Meeus & Raaijmakers 1995: 158-9; Morelli 1983: 187; Rochat & Modigliani 2000: 104). I reply that any such obligation would be substantially

weaker than the obligation to avoid "shocking" the screaming learner (Milgram 1983: 191). (v) The participants were "morally lucky": no shocks were actually administered. I reply that the absence of shocks shows at most that nonsuspicious obedience was less deplorable than it would have been in the presence of shocks, not that it failed to be deplorable: moral luck can get you only so far.[19] Now one might agree with my *individual* replies to the above mitigating factors but claim that the *cumulative* force of these factors (cf. Blass 2000: 43-4) made nonsuspicious obedience nondeplorable. In reply consider a hypothetical variant of Milgram's experiment in which shocks are actually administered. (The confederate may scream because the shocks are really painful, but due to soundproofing the participant hears only the prerecorded protests.) I hope it will be conceded that in such a hypothetical variant nonsuspicious obedience is deplorable *despite* the cumulative force of the above mitigating factors. But this concession is all I need for Q3: given that such a hypothetical experiment is indistinguishable from the actual one as far as the participants are concerned, everyone who would nonsuspiciously obey in the actual would nonsuspiciously obey in such a hypothetical experiment.

(II) Having completed my defense of the claim that most *participants* behaved deplorably, I turn now to the claim that most *people* would be nonsuspiciously obedient. The latter claim is based on two considerations. First, Milgram's participants varied widely in age (20-50) and came from all walks of life: they included "postal clerks, high school teachers, salesmen, engineers, and laborers" (Milgram 1974: 16).[20] Second, high

---

[19] One might object that in some cases of moral luck the behavior fails to be deplorable: consider a truck driver who fails to have his brakes checked but luckily does *not* run over a child (cf. Nagel 1976/1979: 28-9). I reply that such a case of negligence is not appropriately analogous to the nonsuspicious participants' behavior, because they believed they were causing pain *intentionally* (even if unwillingly: footnote 11). A somewhat more appropriate analogy would be with an unsuccessful (and reluctant) attempt at homicide, which is normally deplorable—even if less deplorable than a successful (and nonreluctant) attempt.

[20] Darley (1995: 128-9) argues that, because Milgram's participants were volunteers, they may have been more likely than nonparticipants to value scientific experiments and thus to obey. I reply that only 17% of the participants in follow-up studies mentioned curiosity about psychology experiments as their principal reason for coming to the laboratory (Milgram 1974: 170). Patten (1977b: 435-7) argues that, because vol-

obedience rates were obtained in many variants of Milgram's experiment, conducted both by Milgram at Yale and by others in several countries: Australia (Kilham & Mann 1974), Austria (Schurz 1985), Germany (Mantell 1971; Mantell & Panzarella 1976), Italy (Ancona & Pareyson 1968; cf. Blass 1992: 304), Jordan (a "non-Western" country: Shanab & Yahya 1977, 1978), South Africa (Edwards, Franks, Friedgood, Lobban, & Mackay 1969; cf. Blass 2000: 48, 58-9), Spain (Miranda, Caballero, Gomez, & Zamorano 1981), UK (Burley & McGuinness 1977), and USA (e.g., Bok & Warren 1972; Constanzo 1976; Powers & Geen 1972; Rosenhan 1969: 141-3; cf. Shalala 1974).[21] Nine of the above studies included both male and female participants, but eight of these nine studies found no statistically significant sex difference in obedience rates (Blass 2000: 47-50). I conclude that most people would behave deplorably if they participated in Milgram's experiment.[22] (In response to the claim that Milgram's experimental situation is too "artificial", in §5.2 I describe some more naturalistic related studies.)

---

unteers have a higher need for social approval than nonvolunteers (R. Rosenthal 1965: 394-5; Rosenthal & Rosnow 1975: 40-4; Rosnow 1993: 425; Rosnow & Rosenthal 1976: 99), Milgram's participants may have been more likely than nonparticipants to obey. Pigden & Gillet (1996: 237-9) reply that, if a sufficiently high percentage of people are potential volunteers, and given that the difference in need for social approval between volunteers and nonvolunteers is small, only a slight downward revision may be needed in the percentage of people who would obey.

[21] For reviews see Blass 1991, 1992, 1999, 2000; Miller 1986: chap. 4; Smith & Bond 1993: 19-21. In some variants the percentage of obedient participants was much lower than 80%, but these variants correspond to situations significantly different from the situation of the variant I described above; e.g., "only" 47.5% of participants were obedient (30% were fully obedient) when they had to physically force the learner's hand onto a "shock plate".

[22] One might object that nowadays most people have heard of Milgram's experiments and thus would doubt the reality of the putative shocks (cf. Darley 1995: 142). I reply that in a recent study by Blass (1996a) about 63% of 41 students in introductory psychology at a U.S. university denied any familiarity with the obedience experiments; presumably the percentage among the public (or in many other countries) would be much higher. One might similarly object that nowadays most people know that deception in social psychological experiments is widespread and thus would doubt the reality of the putative shocks (cf. Berkowitz & Donnerstein 1982: 251). I reply that a participant may believe that deception is involved but fail to locate the deception at the reality of the putative shocks; e.g., a participant may believe that the purpose of the experiment is not to investigate the effect of punishment on learning, but rather (as an actual participant said) to "test the effects on the teacher of being in an essentially sadistic role" (Milgram 1974: 53; cf. Shelton 1982: 21). Moreover, a review by Blass (2000: 51) found that obedience rates have not significantly changed over time.

2.2.2. The Stanford Prison Experiment

I turn now to the Stanford Prison Experiment (SPE), conducted by Zimbardo, Haney, Banks, and Jaffe (1973).[23] A newspaper advertisement offering $15 per day for a "psychological study of prison life" elicited more than 70 responses. Twenty-four presumably "emotionally stable, physically healthy, mature, law-abiding" participants were selected. Each participant signed a contract making explicit that those who would be selected to role-play prisoners should expect to have some of their basic civil rights suspended during their imprisonment. On a random basis, twelve participants were selected to role-play guards and twelve to role-play prisoners.[24] The guards attended an orientation meeting in which they were intentionally given only minimal guidelines: their task was to "maintain the reasonable degree of order within the prison necessary for its effective functioning" and to deal with any contingency (e.g., prisoner escape attempts) without ever resorting to physical violence. The prisoners were asked to be available at their residence on day 1 (Sunday, 15 August 1971) but were not told that they would role-play prisoners or given any information about what would happen. On day 1, each prisoner was "arrested" by (real) police, treated like an ordinary suspect (handcuffed, searched, fingerprinted, etc.), placed in a detention cell, and subsequently driven by an experimenter and a guard to the experimental prison (located in the basement of a Stanford University building).

(I) My claim that most guards behaved deplorably is based on the following facts.

> Typically, the guards insulted the prisoners, threatened them, were physically aggressive, used instruments (night sticks, fire extinguishers, etc.) to keep the prisoners

---

[23] See also: Faber 1971; Haney 1976; Haney, Banks, & Zimbardo 1973, 1976; Haney & Zimbardo 1977, 1998; Musen & Zimbardo 1992; White & Zimbardo 1972; Zimbardo 1973a, 1973b, 1975; Zimbardo, Maslach, & Haney 2000; Zimbardo & White 1972.

[24] Initially nine guards (in three eight-hour, three-guard shifts) and nine prisoners participated in the experiment. Later on two stand-by guards and one stand-by prisoner also participated, bringing the number of actual participants to 21.

> in line ... They made the prisoners obey petty, meaningless and often inconsistent rules, forced them to engage in tedious, useless work, such as moving cartons back and forth between closets and picking thorns out of their blankets for hours on end. (The guards had previously dragged the blankets through thorny bushes to create this disagreeable task.) Not only did the prisoners have to sing songs or laugh or refrain from smiling on command; they were also encouraged to curse and vilify each other publicly ... and were repeatedly made to do push-ups, on occasion with a guard stepping on them or a prisoner sitting on them. ... After 10 P.M. lockup, toilet privileges were denied, so prisoners [had] to urinate and defecate in buckets provided by the guards. Sometimes the guards refused permission to have them cleaned out, and this made the prison smell. (Zimbardo et al. 1973: 48, 44, 39.)
>
> [P]ractically all prisoner's rights (even such things as the time and conditions of sleeping and eating) came to be redefined by the guards as "privileges" which were to be earned for obedient behaviour. ... A question by a prisoner as often elicited derogation and aggression as it did a rational answer. Smiling at a joke could be punished in the same way that failing to smile might be. (Haney et al. 1973: 94, 95; cf. 1976: 175, 173.)
>
> [A guard afterwards said:] "... I was a real crumb. I made them call each other names and clean out the toilets with their bare hands. I practically considered the prisoners cattle ..." ... [Another guard] kept a man in the "hole" [an "extremely small" unlit closet used for solitary confinement] for three hours ... and would have left him there all night if one of Zimbardo's assistants had not intervened. (Faber 1971: 83, 82.)
>
> [The] tone became increasingly ugly as guards ... invented new activities to demean the prisoners, mostly by having them enact rituals with a sexual, homophobic character. ... [The guards'] boredom drove them to ever more degrading abuse of the prisoners, ever more pornographic. (Zimbardo & White 1972: 75.)

"When questioned after the study about their persistent affrontive and harassing behaviour in the face of prisoner emotional trauma, most guards replied that they were 'just playing the role' of a tough guard" (Haney et al. 1973: 92-3). One might thus object that most guards did not behave deplorably. Actually there are two possible objections here.

First, some guards may have been playing a role in the sense of doing their job: they had freely signed a contract and thus they felt an obligation to comply with the experimenters' expectations. It is true that the experimenters formulated only minimal guidelines, but they did say that (a) they wanted to "simulate a prison environment within the limits imposed by pragmatic and ethical considerations" (Haney et al. 1973: 74) and that (b) the guards' task was to maintain order within the prison. From the first statement the guards may have inferred that they were expected to behave much like real

prison guards, namely oppressively (Banuazizi & Movahedi 1975),[25] and the second statement is relevant because the harassment of the prisoners by the guards started as a reaction to a rebellion by the prisoners which erupted on the morning of day 2 (Monday). In reply note first that the rebellion was quickly quashed, whereas the harassment "steadily escalated from day to day although prisoner resistance—its original justification—declined and dissolved" (Zimbardo 1975: 49). Moreover, many guards did not harass the prisoners unwillingly: they reported "being delighted in the new-found power and control they exercised and sorry to see it relinquished at the end of the study" (Zimbardo et al. 1973: 49).[26] In addition, "[m]ost of the worst prisoner treatment came on night shifts and other occasions when the guards thought they could avoid the surveillance and interference of the research team" (Haney & Zimbardo 1998: 709), "who were thought to be too soft on the prisoners" (Haney et al. 1973: 92).[27] Finally, any obligation to comply with the experimenters' expectations would be weaker than the obligation to avoid harassing the prisoners.

Second, some guards may have been playing a role in the sense of viewing the experiment much like a game. But the experiment was no game to the prisoners, who were at the mercy of the guards for going to the toilet, drinking a glass of water, or brushing their teeth: all these were "privileged activities requiring permission and neces-

---

[25] Such an inference may have been facilitated by the fact that the experimenters took care to *deindividuate* (Zimbardo 1970) the prisoners, who had to wear numbered smocks, cover their hair with nylon stockings made into caps, address each other only by ID number, and so on.

[26] Cf.: "most of the guards seemed to be distressed by the decision to stop the experiment [prematurely, on the morning of day 6 (Friday)] and it appeared to us that ... they now enjoyed the extreme control and power which they exercised and were reluctant to give it up. ... Many of the guards showed in their behaviour and revealed in post-experimental statements that this sense of power was exhilarating" (Haney et al. 1973: 81, 94).

[27] Cf. DeJong 1975: 1014. Cf. also the following excerpt from a guard's diary: "I am surprised and angry that the psychologist rebukes me for handcuffing and blindfolding the prisoner before leaving the office, and I resentfully reply that it is both necessary security and *my* business anyway" (Haney & Zimbardo 1977: 208; cf. Zimbardo et al. 1973: 53; Zimbardo et al. 2000: 226).

sitating a prior show of good behaviour" (Haney et al. 1973: 96). As a prisoner afterwards put it: "it was a prison to me, it *still* is a prison to me, I don't regard it as an experiment or a simulation. It was just a prison that was run by psychologists instead of run by the state".[28] As early as day 2 a prisoner was released because he exhibited "extreme depression, disorganized thinking, uncontrollable crying and fits of rage" (Zimbardo et al. 1973: 48). During the next three days three more prisoners exhibited similar symptoms and were also released. Some guards thought that these prisoners were faking (Haney & Zimbardo 1977: 209), but what about a prisoner who developed a "psychosomatic rash" when his parole appeal was rejected?[29] The blindness of some guards to the prisoners' suffering was presumably self-serving and does not adequately excuse those guards' behavior.[30]

One might also object that not all guards behaved alike. It is true that "about a third of them were so consistently hostile and degrading as to be described sadistic. They appeared to take pleasure in the prisoners' suffering" (Zimbardo 1975: 46). But other guards were "tough but fair ('played by the rules'), ... while a few [or "several": Zimbardo 1973a: 154; Zimbardo & White 1972: 70] were passive and rarely instigated any coercive control" (Haney et al.: 1973: 81): "they occasionally did little favors for the

---

[28] Haney et al. 1973: 88; Musen & Zimbardo 1992; White & Zimbardo 1972; Zimbardo et al. 2000: 201, 218; Zimbardo & White 1972: 77. Cf. Doyle (1975: 1013): "In a sense, Zimbardo ... did not simulate a real prison, but created a special kind of prison." The seriousness of the prison to the prisoners is also attested by the fact that when they were alone in their cells with each other "almost all (a full 90%) of what they talked about was directly related to immediate prison conditions, that is, food, privileges, punishment, guard harassment, etc." (Haney et al. 1973: 92; cf. 86; Zimbardo et al. 1973: 46).

[29] Zimbardo et al. 1973: 48. Cf. Haney et al. 1973: 81; DeJong 1975: 1014. Contrast Movahedi & Banuazizi 975: 1017.

[30] Another objection to my claim that most guards behaved deplorably can be derived from a guard's statement that he viewed his behavior as degrading but not as really harmful (Musen & Zimbardo 1992; White & Zimbardo 1972). But it seems that humiliating and harassing do constitute harming. Maybe they did not cause *permanent* harm (Haney et al. 1973: 88; Zimbardo 1973b: 249, 254; Zimbardo et al. 1973: 58, 60), but this shows at most that they were less deplorable than they would have been in the presence of permanent harm, not that they failed to be deplorable.

prisoners, were reluctant to punish them, and avoided situations where prisoners were being harassed" (Zimbardo et al. 1973: 49). I reply that *every* guard "behaved at one time or other in abusive, dehumanizing ways" (Zimbardo 1975: 45), and that "even those 'good' guards ... respected the implicit norm of *never* contradicting or even interfering with an action of a more hostile guard on their shift" (Haney et al. 1973: 94).[31]

(II) Two considerations support my claim that most *people* would behave much like those *participants* in SPE who role-played guards did. First, although the participants in SPE did not form a representative sample of people in general (they were middle-class white male college students aged 17-30), they were arguably *less* likely than people in general to behave deplorably: they had been screened (by means of an extensive questionnaire and an interview) for anti-social behavior and emotional instability, and they were "seemingly gentle and caring young men, some of whom had described themselves as pacifists or Vietnam War 'doves' " (Haney & Zimbardo 1998: 709). Second, although we lack replications of SPE,[32] a precursor of SPE was carried out in the spring of 1971 by a group of (both male and female) undergraduates who had been assigned in one of Zimbardo's courses the project of studying prison life (and who, incidentally, "belonged to a dormitory house plan which was dedicated to nonviolence"). The results were apparently similar to those of SPE: "the mock guards dehumanized the

---

[31] Cf.: "Still, the behavior of these good guards seemed more motivated by a desire to be liked by everyone in the system than by a concern for the inmates' welfare. No guard ever intervened in any direct way on behalf of the prisoners, ever interfered with the orders of the cruelest guards or ever openly complained about the subhuman quality of life that characterized this prison" (Zimbardo et al. 1973: 49; cf. Evans 1980: 207; Sabini & Silver 1982: 51, 82; Zimbardo 1973a: 154; Zimbardo et al. 2000: 203).

[32] Close to a replication comes a simulated-prison study at the University of New South Wales. The guards' behavior was "less extreme than the behaviour of the Stanford subjects", but there were "important procedural differences between the two experiments" and "the participants in the U.N.S.W. experiment were subjected to much tighter behavioural constraints than the participants in the Stanford Study"; in particular, harassment by the guards was "prohibited" (Lovibond, Mithiran, & Adams 1979: 283-4; cf. Lovibond & Adams 1979; Morgan 1979). See also Orlando 1973 and Doyle 1975 for role-playing studies (not about prison environments) structurally similar to SPE.

mock prisoners in a variety of ways" (Zimbardo 1975: 37). It seems reasonable to conclude that most people would behave deplorably if they role-played guards in SPE.

2.2.3. The seizure experiments

Intrigued by the inactivity of thirty-eight witnesses to the murder of Kitty Genovese (A. M. Rosenthal 1964), Darley and Latané (1968) studied experimentally people's reactions to a simulated emergency—an epileptic seizure.[33] The participants were undergraduate students who had been recruited for an unspecified experiment as part of a class requirement. Each participant was given a pair of headphones with an attached microphone, was told to listen for instructions, and was left alone in a small room. The experimenter stated over the intercom that a discussion about personal problems associated with college life was to follow, and that the speakers had been placed in individual rooms to preserve anonymity and to avoid embarrassment. The experimenter also stated that, to avoid influencing the discussion, he would not be listening but would get the speakers' reactions later; the discussion would be regulated by a mechanical switching device which would turn successive speakers' microphones on and off about every two minutes, so that only one speaker's microphone would be on at any given time. There were in fact no other speakers, but the participant heard tape-recorded voices. First was heard the voice of the future "victim", who stated that he was having adjustment problems and hesitantly mentioned that he was prone to seizures. Next were successively heard the voices of four other people if the participant was in the *six-person condition*; no voice other than the victim's was heard by the participant in the *two-person condition*. Finally the participant spoke. The second round started again with the voice of the victim, who grew increasingly loud and incoherent, said that he was having

---

[33] See also: Evans 1980: 216-8; Hunt 1990: 132-5; Latané & Darley 1969: 261-5, 1970a: 22-5, 1970b: chap. 11, 1976: 14-7.

a seizure, repeatedly asked for help, choked, and finally broke down (and then remained silent) about 70 seconds after the beginning of his speech.

(I) About 62% of the 13 participants in the six-person condition were *unresponsive*: half a minute after the victim's breakdown they had not left their rooms. One might argue that this lack of responsiveness was largely due to suspicion about the genuineness of the emergency: in reply to a questionnaire asking the participants to check the thoughts which crossed their minds when they heard the victim call for help, about 31% of 65 participants checked "I thought it must be some sort of fake".[34] I have two replies. First, many (behaviorally) unresponsive participants showed signs of conflict rather than apathy (and were thus in a sense *emotionally* responsive): "they often had trembling hands and sweating palms. If anything, they seemed more emotionally aroused than did the subjects who reported the emergency".[35] Second, a full 12 of the 13 participants in the *two*-person condition were responsive,[36] so apparently they found the simulated sei-

---

[34] This argument goes from the *premise* that 31% of 65 participants checked the thought that the seizure was a fake, through the *lemma* that suspicion was widespread, to the *conclusion* that unresponsiveness was largely due to suspicion. In the text I go on to give two reasons to doubt the lemma; here are two reasons to doubt the inference from the premise to the lemma. First, the questionnaire was administered *after* the deception was explained to the participants. Second, the participants were not presented with an *open*-ended question about the thoughts which crossed their minds; the participants were rather asked to *check* thoughts from a 15-item checklist. Such a *closed* question format can lead to biases: in an experiment by Schuman and Scott (1987), "when a widely used open-ended question about 'the most important problem facing this country today' was converted into a closed question listing four specific problems, the listed responses rose dramatically ('quality of public schools' increased from 1 percent to 32 percent), while almost none of the common responses to the open question (e.g., 'unemployment') were offered much despite the encouragement for 'other' answers" (Schwarz, Groves, & Schuman 1998: 160).

[35] Darley & Latané 1968: 382; Latané & Darley 1969: 264, 1970a: 24, 1970b: 100, 1976: 16. (These reactions don't exonerate the participants; cf. footnote 11.) Darley and Latané continue: "Why, then, didn't they respond? It is our impression that nonintervening subjects had not decided *not* to respond. Rather they were still in a state of indecision and conflict concerning whether to respond or not." In the six-person condition about 23% of the 13 participants did leave their rooms more than half a minute after the victim's breakdown, but I submit that this was too late to excuse these participants.

[36] The participants who left their rooms saw an experimental assistant and reported the emergency to the assistant. As Harris and Robinson (1973: 8) note, it cannot be ruled out that these participants were not really "responsive": maybe they were attempting to leave the building and would not have reported the emergency if the assistant had not been immediately visible. But in a replication by Schwartz and Clausen (1970: 303) in which the assistant was not visible, no participant simply left the building.

zure highly credible.[37] (Or at least, given that even in the two-person condition some participants checked the thought that the seizure was a fake,[38] if this thought crossed their minds it resulted in at most *slight* suspicion.) One might object that a lack of suspicion in the two-person condition does not *entail* such a lack in the six-person condition: maybe the participants in the two conditions perceived the same tape-recording differently.[39] I reply that in such a case it's probably the perceptions in the *six*-person condition which should be considered distorted, given that these perceptions would rationalize inactivity. Such a distortion would not constitute an adequate excuse: *this* kind of suspicious unresponsiveness would be deplorable *if* nonsuspicious unresponsiveness was.[40] I conclude that most participants in the six-person condition behaved deplorably if nonsuspicious unresponsiveness in that condition was deplorable—but was it?

It was: it violated the duty to act so as to assist a person that one has good reason to believe is in need when one has the power and the opportunity to do so at negligible cost to oneself.[41] The existence of such a duty is consistent with the view that "Good

---

[37] Apparently they also believed that the experimenter was indeed not listening, so I include this belief in my understanding of lack of suspicion.

[38] I infer this from Darley and Latané's assertion that "there were no significant differences in number or kinds of thoughts in the different experimental groups" (1968: 381; Latané & Darley 1970b: 99). (Given the small sample sizes, however, a lack of "[statistically] significant" differences is compatible with the existence of nonnegligible numerical differences.)

[39] Evidence for this possibility comes from another experiment (Latané & Darley 1970b: 83-5). The participants overheard a protracted tape-recorded fight in which a child was being beaten by another and was crying for help. Then the participants were asked whether they believed that the fight was real. About 88% of the 8 participants who had been told that there was an adult with the children answered positively, whereas only 25% of the 12 participants who had been told that the children were alone did so. In each condition only one participant attempted to intervene, but apparently the participants in the former condition did not need to rationalize their inactivity by distorting their perceptions because they could shift responsibility to the adult who was supposed to be with the children.

[40] It might be objected that one cannot be responsible for behavior due to an *unintentional* perceptual distortion. I reply that such behavior does show a defect of character and is thus relevant to character evaluations.

[41] Cf. Schedler 1982: 400-1. The existence of such a duty arguably does not entail the existence of a *right* to be assisted when in need (cf. Thomson 1971/1993: 183).

Samaritan" laws should not be enacted: not every moral duty should be enshrined in law.[42] The participants did have the power to assist: not *directly* (they may not have known how to deal with a seizure victim and in any case they did not know in which room the victim was supposed to be), but at least they could *report* the emergency. One might try to find excuses for the participants who violated the above duty: maybe they were concerned not to embarrass the victim, "not to make fools of themselves by overreacting, not to ruin the ongoing experiment by leaving their intercom, and not to destroy the anonymous nature of the situation which the experimenter had earlier stressed as important".[43] Three points in reply. First, reporting the emergency would neither unduly embarrass the victim nor count as overreacting, because the victim had asked for help *six times*. Second, given the victim's breakdown and subsequent silence, the experiment was already ruined as far as the participants knew. Third, a desire to preserve anonymity, though admittedly legitimate, should have carried little weight when pitted against the victim's apparent need.[44]

One might grant that nonsuspicious unresponsiveness in the *two*-person condition was deplorable but claim that participants in the *six*-person condition had a special excuse: they believed that four other people were listening. But how would this belief provide an excuse? Certainly not via a callous "why me" reasoning: "four other people can

---

[42] The phrase "Good Samaritan laws" sometimes refers to laws which require people to assist others in emergencies and other times refers to laws which confer immunity from liability on people who assist others in emergencies; I adopt the former use.

[43] Darley & Latané 1968: 382; cf. Latané & Darley 1969: 264, 1970a: 25, 1970b: 101, 1976: 16. The statement that they did not want to embarrass the victim was made by three nonintervening participants in *another* bystander intervention experiment (Latané & Rodin 1969: 197; Latané & Darley 1969: 256, 1970a: 20, 1970b: 65, 1976: 11).

[44] One might also argue that the unresponsive participants were adequately excused because they were temporarily paralyzed (they would have acted later) or morally lucky (there was in fact no seizure). I reply to the former point in footnote 35; for my reply to the latter see footnote 19 and the corresponding text.

help, so why should I do so?"[45] Maybe via a "redundancy" reasoning: "someone else is probably helping, so my own help would be superfluous". (Either reasoning was probably at most implicit: all participants in the six-person condition denied that their awareness of the presence of other people made a difference to their own behavior.[46]) But what would prompt a redundancy reasoning, given the participants' ignorance of what the other listeners were doing? Maybe a "statistical" reasoning: "the more bystanders there are, the more likely it is that at least one will help". This reasoning is not foolproof: maybe the more bystanders there are, the more likely each of them is to assume that someone else is helping, and the less likely it is that the victim will get help (cf. Latané & Nida 1981: 321-2; Latané, Nida, & Wilson 1981: 307). But even if the reasoning worked, only a *large* number of bystanders could have made the probability that the victim was being helped so high as to excuse unresponsiveness:[47] with only four other people present, the risk that the victim was not being helped was nonnegligible and should have made the participants disregard the possibility that their own help would be super-

---

[45] According to Sabini and Silver, "one has a responsibility and standing to intervene on another's behalf as a consequence of the *contrast* between the relationship that an actor has to the party in distress and the relationships that others have to him. These relationships may be spatial (the closest person to another should help), ... or familial (family members have particular rights and responsibilities to become involved" (1982: 43; cf. Kamm 1999). But even if one grants that in the presence of some asymmetries some bystanders have no duty to intervene (e.g., only the bystander who is a relative of the victim has such a duty), it doesn't follow that in the absence of asymmetries (as in Darley and Latané's experiment) no bystander has a duty to intervene.

[46] Darley & Latané 1968: 381; Latané & Darley 1970b: 100. Cf.: "We asked the question every way we knew how: subtly, directly, tactfully, bluntly, and the answer was always the same. Subjects had been aware of the presence of other bystanders in the appropriate conditions, but they did not feel that they had been influenced in any way by the bystanders' presence" (Latané & Darley 1976: 17; cf. 1969: 265). On the other hand, in a replication by Schwartz and Clausen (1970: 307) 50% of all participants answered "yes" when asked: "Do you think that the presence of (fact there were no) others in your group influenced how you reacted to the emergency?"

[47] I say "*could* have made"—rather than *would*—because I don't want to take a stand on whether the believed presence of even (e.g.) a million bystanders would have excused unresponsiveness.

fluous.[48]

(II) Darley and Latané's (1968) experiment was replicated several times (Schwartz & Clausen 1970; Horowitz 1971; Harris & Robinson 1973; Schwartz & Gottlieb 1980). For my purposes the reports of most replications are less than ideally detailed, but on the whole I take the replications to support my claim that most people would be nonsuspiciously unresponsive in the six-person condition of Darley and Latané's experiment.[49]

This completes my defense of the claim that (Q3) there are many situations in each of which most people (would) behave deplorably.[50]

---

[48] One might also try to find a special excuse for nonsuspicious unresponsiveness in the six-person condition by appealing to the principle that, when N people are collectively responsible (cf. Schedler 1982) for an event and their roles are symmetric, then each of them is responsible only to degree $1/N$ for the event. I don't need to take a stand on this principle because it does not provide an adequate excuse: even if in an emergency each bystander is only to a small degree responsible for the fact that the victim receives no help, each bystander can still be fully responsible for the fact that she herself does not help—and it's the latter fact which corresponds to deplorable behavior for my purposes.

[49] (1) Harris and Robinson (1973)—who used an asthma attack rather than a seizure—report results only on speed of intervention, not on percentages of intervening participants. (2) Schwartz and Gottlieb (1980) report very high percentages of intervening participants, but in their experiment the participants could *see* as well as hear the victim and could report the emergency by telephoning the experimenter and thus without leaving their rooms. (3) Horowitz (1971) reports that in the two-person condition 55% of males drawn from social fraternity groups and again 55% of males drawn from an on campus community service organization left their rooms by the end of the experiment, whereas the corresponding percentages in a five-person condition were 65% and 20% respectively. (4) Finally, Schwartz and Clausen (1970) report that 100% of females in the two-person condition but 33% of females in the six-person condition left their rooms *by the end of the seizure*, whereas 92% of females and 72% of males in an amalgam of two-person conditions, in contrast to 69% of females and 74% of males in an amalgam of six-person conditions left their rooms *by the end of the experiment*. (Darley and Latané, whose participants in the two- and six-person conditions were all female, report that 85% of participants in the two- and 31% of participants in the six-person condition left their rooms by the end of the seizure, whereas 100% of participants in the two- and 62% of participants in the six-person condition left their rooms by the end of the experiment.) It seems thus reasonable to infer that *half a minute* after the end of the seizure most males from either group in Horowitz's five-person condition, as well as most females and most males in Schwartz and Clausen's six-person conditions had not left their rooms. (I defined *unresponsiveness* as the failure to leave one's room *half a minute* after the end of the seizure because I think that (a) leaving one's room *later*—i.e., by the end of the experiment, six minutes after the start of the seizure—is not excusable (cf. footnote 35) and that (b) failing to leave one's room *by the end of the seizure* need not be deplorable.)

[50] There are many other experiments in each of which most participants in some experimental condition fail to help although help is easy (e.g.: Clark & Word 1972; Darley, Teger, & Lewis 1973; Gaertner 1975; Gaertner & Dovidio 1977; Latané & Rodin 1969; A. S. Ross 1971; Ross & Braband 1973; Rutkowski,

2.3. Situations in which most people would behave admirably (Q4)

Those who are tempted to derive a bleak picture of human nature from Milgram's and Zimbardo's experiments[51] would do well to remember that these experiments correspond to a biased sample of situations, selected precisely to exemplify deplorable behavior. A more complex picture emerges when one combines these experiments with those I will examine to support the claim that (Q4) there are many situations in each of which most people (would) behave admirably: (1) the theft experiments (in which most participants who have agreed to watch someone's things stop a confederate from "stealing" them), (2) the electrocution experiments (in which most participants help an apparently electrocuted confederate at the risk of being electrocuted themselves), and (3) the rape experiments (in which most participants try to stop a simulated rape).[52] These

---

Gruder, & Romer 1983; Shotland & Heinold 1985; Smith, Smythe, & Lien 1972; Smith, Vanderbilt, & Callen 1973; Staub 1970, 1971, 1974; Yakimovich & Saltz 1971; for reviews see: Latané & Nida 1981; Latané, Nida, & Wilson 1981; Piliavin, Dovidio, Gaertner, & Clark 1981). I am not describing these experiments because I do not think that in any of them the failure to help is *clearly* deplorable (e.g., because the simulated emergency is ambiguous or insufficiently severe, or because participants are influenced by passive confederates).

[51] Ingram, for example, talks about "Milgram's evidence for a Hobbesian view of human nature" (1979: 529; cf. Patten 1977b: 440; Pigden & Gillet 1996: 240-1). One might be tempted to derive such a bleak picture because of what is variously called "the fundamental attribution error" (L. D. Ross 1977), "lay dispositionalism" (Ross & Nisbett 1991), or "the correspondence bias" (Gilbert & Malone 1995) (although Choi, Nisbett, and Norenzayan 1999 distinguish between these three terms), namely the tendency to overestimate the role of personality and underestimate the role of situational factors when explaining behavior. This tendency was exhibited, for example, in an experiment by Safer (1980): participants who watched Milgram's (1965a) film *over*estimated (in comparison with both "naive" participants and the actual result of 2.5%) the percentage of those who administer the maximum shock in a variant of Milgram's (1974: 70-2) experiment in which people are instructed to *choose* shock levels. (For further reactions of people to Milgram's experiments see: Bierbrauer 1979; Blass 1996b; Miller 1986: 22-33; Miller, Gillen, Schenker, & Radlove 1974; Sabini & Silver 1983.)

[52] An even more complex picture emerges when one also takes into account variants of Milgram's experiment in which (1) most participants *don't* harm the learner and (2) most participants (think they) harm *themselves*. (1a) It has been consistently found that, when participants are instructed to *choose* shock levels, very few administer the highest shock (for a review see Blass 1992: 299-302). (1b) Only 4 out of 40 participants were fully obedient when they witnessed two confederates disobey (Milgram 1965c, 1974: 116-21). (2a) In an experiment by Martin et al. (1976), 21 out of 39 participants thought they were administering to their own ears ultra-high frequency sounds likely to result in severe hearing loss. (2b) In an experiment by Kudirka (1965), 14 out of 19 participants complied with the experimenter's request to eat 18 "extremely bitter and distasteful" quinine-soaked crackers.

three kinds of experiments stand in marked contrast to the seizure experiments (§2.2.3): it is intriguing that in some bystander intervention experiments most participants try to help although help is risky while in other such experiments most participants fail to help although help is easy.[53] Such large variations in average behavior in response to apparently even negligible differences between experimental conditions are a recurrent theme in the psychological literature on helping behavior. Darley and Batson (1973), for example, report that 63% of *nonhurried* participants helped a coughing and groaning confederate who was sitting slumped in a doorway, while only 10% of *hurried* participants helped.[54] Isen and Levin (1972), as another example, report that 87.5% of those participants who had just found a dime in the coin return slot of a public telephone helped a confederate who "accidentally" dropped a folder full of papers, while only 4% of those participants who had found no coin helped.[55] I will not go into details because I think that in these experiments failing to help is not sufficiently blameworthy to count as deplorable and helping is not sufficiently praiseworthy to count as admirable.[56] So I turn now to the experiments I will use to support Q4; for each of them I will argue that (I)

---

[53] This is to say that human behavior is highly complex, not that it has no rhyme or reason. Social psychologists have developed elaborate models to explain and predict helping behavior (Batson 1991; Dovidio et al. 1991; Latané, Nida, & Wilson 1981; Piliavin et al. 1981, 1982).

[54] Cf. Batson 1976; Campbell 1999; Doris 1996, in press. The complexity of human behavior is further illustrated by the fact that, in a replication of Darley and Batson's (1973) experiment (Batson et al. 1978), 70% of hurried participants who had been told that their participation was of low importance to the experimenter did help.

[55] Levin and Isen (1975) report two successful replications. (They also say that postage for a letter was 8¢ back then, so a dime was a less negligible amount than it is now.) For related experiments see: Aderman & Berkowitz 1970; Berkowitz & Connor 1966; Isen 1970; Isen, Clark, & Schwartz 1976; Isen, Horn, & Rosenhan 1973; Isen, Shalker, Clark, & Karp 1978; Isen & Simmonds 1978. For reviews see: Carlson, Charlin, & Miller 1988; Carlson & Miller 1987; Salovey, Mayer, & Rosenhan 1991; Schaller & Cialdini 1990.

[56] Another group of experiments in which I think that help does not count as admirable are reported by Batson (1991; cf. Grant 1997): participants who were watching a confederate "suffer" from the administration of electric shocks agreed to take the confederate's place. This behavior need not be admirable because the participants were led to believe that the confederate's extreme reaction to the shocks was atypical and that they would not find the shocks as unpleasant if they chose to take the confederate's place.

most *participants* behaved admirably and that (II) most *people* would behave as those participants who behaved admirably did.

2.3.1. The theft experiments

Theft on the beach. On a summer weekday you are relaxing on the beach. A woman in her middle 20s, dressed in usual beach attire, places her blanket close to yours, turns on a portable radio (tuned to a local rock station) at a fairly high volume, and reclines for a couple of minutes. Then she leaves her blanket and addresses you: "Excuse me, I'm going up to the boardwalk for a few minutes ... would you watch my things?" You agree to do so. A few minutes later, a tall man in his middle 20s walks up to the woman's blanket, picks up the radio (which is still playing), and quickly walks away. What do you do?

Theft in the restaurant. On a spring weekday you are dining alone at an Automat cafeteria in midtown Manhattan. A well-dressed woman in her early 20s sits at your table. After a few minutes she lights a cigarette and then, pointing to her suitcase on the table, asks you: "Excuse me ... may I leave this here for a few minutes?" You respond affirmatively and she walks away. A few minutes later, a man in his early 20s approaches the table, picks up the suitcase, and quickly walks away. How do you react?

(I) All 10 (unwitting) participants in the beach experiment and all 8 participants in the restaurant experiment ran after the "thief" (in fact a confederate) and stopped him (Moriarty 1975).[57] Did they behave admirably? First, it's hard to ascribe an ulterior motive to them.[58] (A speculation that they might have wanted to befriend the female

---

[57] In the beach experiment the sampling units were beach blankets rather than individuals. "Where the blanket was occupied by more than one person, the victim's dialogue was addressed to the person most attentive but was generally overheard by the other(s) present" (Moriarty 1975: 371).

[58] In the *no-commitment condition* (when the victim asked the participant for a light rather than a commitment) only 4 out of 36 participants in the beach experiment and only 1 out of 8 participants in the restaurant experiment stopped the thief. So it might be suggested that the participants who stopped the thief in

"victim" won't do: in the beach experiment 9 out of 10 participants intervened when the victim was male and the thief female.[59]) Second, I submit that they went beyond the call of duty: they risked a physical confrontation with the thief. It might be objected that they had agreed to watch the victim's belongings and were thus under an obligation to intervene. True,[60] but they could have intervened by just *shouting* at the thief from a safe distance: the expected cost of a physical confrontation with the thief was high enough to make *stopping* the thief supererogatory. Now one might agree that stopping the thief was praiseworthy but claim that it was not *sufficiently* praiseworthy to count as admirable because the danger was not sufficiently severe. Maybe. But there is some evidence that most participants would have stopped the thief even if the danger had been somewhat greater: in another theft experiment with female participants at a corridor of a university building, W. Austin (1979: 2115) found that about the same high percentage (around 65-70%) of participants stopped an "average-sized" and an "extremely large" male thief.

(II) My claim that most people would stop the thief in Moriarty's experiments is based on two considerations. First, Moriarty's participants varied widely in age (from 14 to 70 years) and education ("from elementary school through professional training"). Second, findings similar to Moriarty's were obtained in several experiments. Three of these experiments were conducted at university libraries. (i) Schwarz, Jennings, Petrillo, and Kidd (1980) found that all 10 participants who had agreed to watch a female vic-

---

the commitment condition did so in order to avoid the embarrassment they would otherwise feel later on when facing the victim. This may well be true but would not defeat an ascription of praiseworthiness: similarly to the way in which an action can be blameworthy without being performed for the sake of violating one's duty or of hurting someone (cf. the blameworthiness of *reluctant* nonsuspicious obedience in Milgram's experiment), an action can be praiseworthy without being performed for the sake of exceeding one's duty or of benefiting someone.

[59] "While subjects were firm in detaining the male thief, the treatment accorded the female thief was far more aggressive, subjects being more likely to use physical force. ... 73% [of those who stopped the thief] snatched the radio away from the female thief, while only 17% did so when the thief was male" (Moriarty 1975: 373 n. 4).

[60] Although, given the wording of the victim's request, in the restaurant experiment the commitment to watch the suitcase was *implicit*.

tim's things prevented a male thief from stealing the victim's calculator. (ii) Shaffer, Rogel, and Hendrick (1975) found that 10 of the 12 male and 6 of the 12 female participants who had agreed to watch a victim's things intervened to stop a thief;[61] (iii) they also found in a replication that 5 out of 8 males and 7 out of 8 females intervened.[62] Finally, in a large-scale experiment at a corridor of a university building, W. Austin (1979: 2116-9) found that about 76% of 176 participants (61 out of 88 males and 72 out of 88 females) who had agreed to watch a victim's folders and calculator intervened to stop a thief.[63] All of the above investigators (as well as several others[64]) also consistently found that most participants *failed* to prevent thefts when they were *not* asked to watch the victim's belongings; but this finding poses no threat to my defense of Q4, because I need only the existence of *some* (open list of) situations in which most people would behave admirably.

---

[61] The thief tried to steal the victim's wristwatch (when the victim was male and the thief female) or $20 from the victim's wallet (when the victim was female and the thief male). Among the intervening participants were two males and one female who questioned the thief's motives but did not prevent the theft: they accepted the thief's story that she (or he) had been sent by the victim to pick up the wristwatch (or the money).

[62] These 16 participants were alone (save for the victim or the thief) when the victim's request for a commitment and the attempted theft took place; when the participants were in the presence of a passive confederate, 4 out of 8 males and also 4 out of 8 females intervened.

[63] Interestingly, only about 45% of 176 participants (22 out of 88 males and 58 out of 88 females) intervened to stop the theft of a victim's folders and *books* (rather than *calculator*); similarly, in another experiment W. Austin (1979: 2115-6) found that about 83% of 24 participants intervened to stop the theft of high-value items but only about 17% of 24 participants intervened to stop the theft of low-value items.

[64] DeJong, Marber, & Shaver 1980; Denner 1968; Howard & Crano 1974; Latané & Darley 1970b: 70-3; Schwartz & Gottlieb 1976. (Cf. Sabini & Silver 1982: 42.) Relevant are also experiments in which most participants (unless prompted by a confederate) fail to report staged shoplifting incidents: Bickman 1979; Bickman & Green 1977; Bickman & Rosenbaum 1977; Gelfand, Hartman, Walder, & Page 1973; Latané & Darley 1969: 258-9, 1970b: 74-7, 1976: 12-3; cf. Farrington 1979: 225, 240.

2.3.2. The electrocution experiments

Around 1972 dozens of male Florida State University students were randomly approached on campus and were offered $2 to participate in a single-session validation of a mental ability test to be used by the university administration for evaluating first-year students. Each participant was met in a faculty member's office by a research assistant who led him to a room where he would take the test and then went away. On their way to the room they passed by a laboratory filled with various items of electronic equipment. Numerous high-voltage signs were placed inside and outside the laboratory, and just inside its door a male technician could be seen adjusting an instrument. After the participant completed the test, he passed again by the laboratory on his way out of the building. The technician could be seen on his knees with his back to the door, making repairs on a small switchboard. Suddenly the participant saw a flash of light and heard a dull buzzing sound; the technician stiffened his body, gave out a sharp cry of pain, upset the apparatus and his tools, and collapsed in a prone position on the floor, lying on several wires with one hand resting in the switchboard and the other holding an electronic probe (Clark & Word 1974).

(I) All but one of the participants helped the "technician" (in fact a confederate), and about 71% of those who helped did so *directly*: they separated the technician from the equipment and the wires.[65] In §2.2.3 I said that there is a "duty to act so as to assist a person that one has good reason to believe is in need when one has the power and the

---

[65] The 71% figure is approximate. Clark and Word give the following results (1974: 284):

Participants experience emergency:	Ambiguity of condition		
	Nonambiguous **(described in text)**	*Moderately ambiguous* (technician not seen)	*Highly ambiguous* (technician neither seen nor cries)
Individually	**7 of 8 individuals help**	6 of 8 individuals help	1 of 8 individuals helps
In pairs	**8 of 8 pairs help**	6 of 8 pairs help	3 of 8 pairs help

Clark and Word report that 71% (i.e., 22) of the 31 helpers helped directly and that "direct and indirect help did not interact with the experimental conditions" (1974: 284 n. 4); but given the small sample sizes, the lack of statistically significant differences between the percentages of direct helpers in the three levels of ambiguity does not entail that these three percentages are *very* close to each other.

opportunity to do so at negligible cost to oneself". Not at *severe* cost, however: the participants could have just reported the emergency, so those who helped directly went beyond the call of duty. It's true that many of them said that they had either formal training or experience with electronic equipment; these *competent* direct helpers helped in a safe manner, so one might argue that they did not *perceive* their behavior as dangerous.[66] I have two replies. First, the technician was presumably also competent, so his apparent electrocution gave even competent direct helpers a *reason* to perceive their behavior as dangerous. Maybe they did not *realize* that they had this reason, but arguably their behavior was still admirable if this epistemic failure was due to their hurry to help. Second, about 64% of direct helpers were not competent;[67] many of them even touched the technician with their hands. All but one of those who did so "indicated later that they realized the 'inappropriateness' of their actions, but at the time they acted so quickly that no consideration was given to the possible harm involved" (Clark & Word 1974: 286). It might be argued that such *impulsive* behavior is not admirable: one deserves no credit for a knee-jerk reaction. In reply note first that touching the technician was not quite as automatic as catching a falling vase, pressing a brake, or ducking a punch: the participants had to cover a short distance to get into the laboratory (and maybe also had to drop whatever they were holding). Moreover (and more important), one can deserve credit even for automatic reactions like catching a falling vase[68] (Wright 1974: 45). First, because such

---

[66] Clark and Word say that 79% of the participants (in the three experimental conditions combined) "reported that they interpreted the situation to be dangerous for themselves" (1974: 286); it seems reasonable to assume that the remaining 10 participants were the 10 competent ones.

[67] Clark and Word do not report the 64% figure directly, but I infer it from the facts that there were (a) 22 direct helpers (see footnote 65) and (b) 8 competent direct helpers. The latter fact I infer from Clark and Word's statements that (i) "10 subjects were classified as competent", (ii) "Ninety per cent [i.e., 9] of the competent subjects reacted to the emergency across all conditions", and (iii) "All but one of the competent helpers rendered assistance directly to the victim" (1974: 284).

[68] Or catching a falling person: in an experiment in which a confederate "staggered forward and collapsed" in the New York subway (Piliavin, Rodin, & Piliavin 1969: 291; cf. Piliavin & Piliavin 1972; Piliavin,

reactions are *intentional* and *preventable*: my leg jerks when struck whether I want it to move or not, but if I hate the falling vase I'm probably not going to catch it. Second, because in many emergencies the optimal reaction is impulsive: stopping to deliberate wastes precious time.

(II) The generalizability of the above results from most *participants* to most *people* is supported by the fact that a virtually identical experiment yielded similar results (Clark & Word 1974: 280-3). On the other hand, we lack replications with participants other than male college students.

2.3.3. The rape experiments

On a late evening you are walking towards a secluded but well-lit campus parking lot. A woman is walking ahead of you, with her back to you. Suddenly a tall man emerges from the bushes and grabs the woman roughly, putting one hand over her mouth and the other around her waist. They struggle; the woman's books fall to the ground and scatter. The woman emits a muffled scream, then a clear scream, another muffled scream, and a yell: "Help! Help! Please help me! You bastard! Rape! Rape!" The man drags her away while she keeps struggling and protesting.

You are at a point where three paths meet, all of them leading to the parking lot. The path straight ahead is shortest but goes through the scene of the incident. The path to your left is moderately longer. The path to your right is considerably longer but a campus security officer can be seen on it, apparently writing parking tickets; he seems too distant to have seen or heard anything. Which path do you choose?

Piliavin, & Rodin 1975), "in many cases bystanders jumped up immediately and intervened, on occasion catching the victim before he hit the floor" (Piliavin et al. 1981: 161, cf. 16).

(I) Half of the 40 (unwitting) participants chose to continue straight ahead.[69] They were stopped and debriefed before they reached the scene of the (staged[70]) incident, so it is not *certain* that they would have intervened: possibly they were just hurrying to reach the parking lot, hoping to avoid encountering the "rapist". But a hidden observer judged that they were proceeding with the intention of helping, given "what they were saying and/or their body language" (Harari, Harari, & White 1985: 656). If so, then they behaved admirably: they could just have alerted the security officer, but they went beyond the call of duty in risking a physical confrontation with the rapist.

(II) Support for the generalizability of the above result comes from two other experiments. (i) Piliavin et al. (1981: 93, 133, 164, 172) report that, in an experiment by Anderson (1974), all 31 participants "rushed out of the experimental room to intervene" when they heard a stranger who "sounded very tough" attempt to rape the experimenter. (ii) In an experiment by Shotland and Straw (1976), when the participants (who were undergraduate students and received course credit for participating) arrived (individually) at the appointed room, they found a note saying that the experimenter would come later and asking them to fill out a questionnaire in between. Five minutes later an incident was staged in the corridor, about 16 meters away from the room. A man physically attacked a woman, violently shaking her while she struggled. She repeatedly exclaimed "I don't know you" and emitted loud piercing shrieks interspersed with pleas to "get away from

---

[69] Six further participants alerted the security officer. In another experimental condition, in which participants experienced the emergency in groups of three to five men who happened to walk close to each other, 28 out of 40 groups continued straight ahead and 6 alerted the security officer.

[70] The woman was a drama major with acting experience. The security officer was real: the experiment was supervised by the campus police. There were also three hidden observers furnished with binoculars and walkie-talkies.

me".[71] Rather than using the university phone in the room, half of 10 male participants intervened directly by approaching the attacker (who fled as a result). More importantly (given that all participants in the experiments by Harari et al. and by Anderson were *male*), half of 10 *female* participants also intervened directly.[72, 73]

This completes my defense of the claim that (Q4) there are many situations in each of which most people (would) behave admirably.

2.4. The validity of the argument from Q3 and Q4 to Q1

I said in §2.1 that the validity of the argument from Q3 and Q4 to the conclusion that (Q1) most people are fragmented is not obvious. Suppose, for example, that in each of three million situations 75% of people behave deplorably. It doesn't follow that each

---

[71] The man and the woman had theatrical training. The attack "was so convincing that on two occasions a faculty member, and then a graduate student, who had both been informed of the experiment and asked to ignore it, nevertheless attempted to intervene" (Shotland & Straw 1976: 992).

[72] No participant used the phone, but one female and two male participants intervened indirectly by alerting other people. Interestingly, in another experimental condition in which the woman repeatedly exclaimed "I don't know why I ever married you" (rather than "I don't know you"), only 2 of 10 male and 1 of 10 female participants intervened directly (one further female participant intervened indirectly by using the phone).

[73] Another experiment, however, yielded different results (Shotland & Stebbins 1980). After completing a (mock) experiment on eye-hand coordination, the participants (again undergraduate students) were sent to an office on a different floor to get an extra-credit card. On the door of that office they found a note saying that its occupant would be back shortly. After a couple of minutes a loud bang was heard as a confederate's leg hit a metal door at the end of a darkened corridor, about 30 meters away from the office. For about 3 seconds a man could be seen, struggling silently with a woman; he shoved her into a room and closed the door. Then a tape-recorded dialogue began: the man tried to rape the woman while she resisted, mostly calmly but occasionally screaming. Only 8 of 40 participants intervened directly by approaching the room (although 19 additional participants intervened indirectly by alerting other people). I find this staged incident less compelling than the previous ones: the woman's protests were much less vehement, and she was silent while she was visible to the participants. Shotland and Stebbins (1980: 524) propose another explanation for the difference in results: a participant who entered the room would *have* to confront the rapist, while a participant in Shotland and Straw's experiment who approached the attacker could hope that the latter would flee along the corridor. [This explanation is supported by a study in which a film of the attack in Shotland and Straw's experiment was shown to 56 students, about 58% of whom said that the man would run away if they intervened (Shotland & Straw 1976: Experiment 4).] Similarly, Harari et al. (1985: 657) suggest that the direct helpers in their own experiment hoped that the rapist would flee as they approached. But even if they did so hope, direct help was still admirable: a nonnegligible risk of physical confrontation with the violent "rapist" remained.

of 75% of people behaves deplorably in all three million situations, because maybe *not the same* people behave deplorably in all situations. Nevertheless, there is a *lower bound* on the percentage of people each of whom behaves deplorably in (e.g.) *more than one third* of the three million situations: at least 62.5%, as it turns out. More generally:

Theorem 2.1. *Consider P people and S_D situations in which on average π_D people (do or would) behave deplorably. Let F_D be the number of people each of whom behaves deplorably in more than σ_D situations ($\sigma_D < S_D$), and let k_D be the average percentage of the S_D situations in which these F_D people behave deplorably. Then:*

$$F_D/P \geq [(\pi_D/P) - (\sigma_D/S_D)]/[k_D - (\sigma_D/S_D)].$$[74]

One can take the lower bound on F_D/P which is given by Theorem 2.1 to be arbitrarily close to π_D/Pk_D because one can take σ_D/S_D to be arbitrarily close to zero: no matter how large σ_D must be to correspond to an open list of situations, given my argument for Q3 (§2.2) one can consider an indefinitely large—and thus an enormously larger than σ_D—number of (counterfactual) situations in which on average π_D people (would) behave deplorably.[75] An estimate of π_D/P is 75%, namely the average of 64%, 100%, and 62% (the approximate percentages of participants who behaved deplorably in Milgram's, Zimbardo's, and Darley and Latané's experiments).[76] An estimate of k_D is

---

[74] *Proof.* Let P_{Ds} be the number of people who behave deplorably in situation s ($s = 1, \ldots, S_D$) and S_{Dp} be the number of situations in which person p behaves deplorably ($p = 1, \ldots, P$). Imagine a matrix whose rows correspond to persons, whose columns correspond to situations, and whose cells have a 'D' exactly if the row-person behaves deplorably in the column-situation. Then the number of D's in column s is P_{Ds}, the number of D's in row p is S_{Dp}, and the total number of D's in the matrix is *both* $\Sigma_s P_{Ds}$ and $\Sigma_p S_{Dp}$: these two sums are equal (1). By definition, $\pi_D = \Sigma_s P_{Ds}/S_D$ (2). Let $\boldsymbol{F}$ be the set of the F_D people each of whom behaves deplorably in more than σ_D situations: $S_{Dp} > \sigma_D$ for every p in $\boldsymbol{F}$, and $S_{Dp} \leq \sigma_D$ for every p in the complementary set $\boldsymbol{F}'$ (which contains P-F_D people). Now $\Sigma_p S_{Dp} = \Sigma_{p \in F} S_{Dp} + \Sigma_{p \in F'} S_{Dp} \leq k_D S_D F_D + \sigma_D(P\text{-}F_D)$ (3). Combining (1), (2), and (3), we get: $S_D\pi_D \leq k_D S_D F_D + \sigma_D(P\text{-}F_D)$, which is equivalent to $F_D \geq (S_D\pi_D\text{-}\sigma_D P)/(k_D S_D\text{-}\sigma_D)$, from which Theorem 2.1 immediately follows. □ (Applied to the numerical example I gave in the text, Theorem 2.1 gives: $F_D/P \geq (75\%\text{-}1/3)/(k_D\text{-}1/3) \geq 62.5\%$ because $k_D \leq 1$.)

[75] There may even be *infinitely* many such situations, but S_D (which is supposed to be finite in Theorem 2.1) need not be the *total* number of such situations.

[76] Recall that in Zimbardo's experiment even the "good" guards behaved deplorably (footnote 31).

88%, namely the midpoint between 75% (the above estimate of π_D/P) and one: k_D is equal to one only in the unlikely case in which *every* person who behaves deplorably in more than σ_D situations behaves deplorably in *all* S_D situations.[77] It follows that an estimate of π_D/Pk_D, and thus an approximate lower bound on F_D/P, is 75%/88% ≅ 85%.

Let F_A be the number of people each of whom behaves *admirably* in more than σ_A of S_A situations ($\sigma_A < S_A$) in which on average π_A people behave admirably. An estimate of π_A/P is 71%, namely the average of 100%, 62.5%, and 50% (the approximate percentages of participants who behaved admirably in the theft, electrocution, and rape experiments).[78] An estimate of k_A (defined analogously to k_D) is 85%, the midpoint between 71% and one. By means of a theorem analogous to Theorem 2.1, it follows that an approximate lower bound on F_A/P is 71%/85% ≅ 84%.

Let finally F be the number of people each of whom behaves deplorably in more than σ_D of the former S_D situations *and* behaves admirably in more than σ_A of the latter S_A situations. It can be shown that $F \geq F_D + F_A - P$,[79] so an approximate lower bound on F/P is 85% + 84% - 100% = 69%: *most* people are fragmented.

This completes my defense of the claim that (Q1) most people are fragmented. I conclude this section with a general remark. I am frequently asked *why* most people are fragmented. I don't know (although one might speculate that being fragmented is evolutionarily adaptive), but why would I need to provide an answer? Maybe the question is

---

[77] The average percentage of the S_D situations in which the P people behave deplorably can be shown to be π_D/P, but one might claim that k_D is still close to one because there are only two kinds of people: those who almost always behave deplorably and those who almost never do so. I reply that there is empirical evidence against this claim: in a series of tests only small percentages of children (almost) always or never cheated (Hartshorne & May 1928: 386).

[78] 62.5% is 5/8: in the electrocution experiment described in §2.3.2, probably 5 (i.e., 71% of 7) of 8 individuals helped directly in the nonambiguous condition (footnote 65).

[79] *Proof.* Let F_{DA} be the number of people each of whom behaves deplorably in more than σ_D of the S_D situations *or* behaves admirably in more than σ_A of the S_A situations. Then $F_{DA} = F_D + F_A - F \leq P$, so $F \geq F_D + F_A - P$. □

motivated by the worry that, even if most people are behaviorally "inconsistent" in the *specific* sense of being fragmented, possibly they are consistent in some other, *deeper* sense. This possibility, however, makes no difference to my reasoning if, as I argue next, the kind of inconsistency which I label 'fragmentation' suffices for indeterminacy.[80]

3. Fragmentation entails indeterminacy (Q2)

3.1. The concept of indeterminacy and an argument for Q2

I call a person *indeterminate* exactly if the person—equivalently, the person's (moral) character—is neither good nor bad nor intermediate; in other words, the person has no *character status*, understood as status on the good/intermediate/bad scale. The claim that a person is indeterminate presupposes neither that our information about the person is imperfect nor that it is perfect: it is a claim not to the effect that we know or believe something about the person, but rather to the effect that the person has a certain property, namely indeterminacy. Indeterminacy is not the property of *having* some peculiar ("indeterminate") character status, but is rather the property of *lacking* character status: it makes no sense in my usage to talk about the character status of an indeterminate person or to say that a person has an "indeterminate character status". (As an analogy, a mathematical sequence is called *divergent* exactly if it has no limit; it makes no sense to talk about the limit of a divergent sequence or to say that a sequence has a "divergent limit".) Although the function which assigns to people their character status is *undefined* for an indeterminate person, to claim that a person is indeterminate is not to claim that our evaluative practice is *silent* on the question of what (if any) is the person's

---

[80] A more worrisome possibility is that there is a hidden consistency in the kinds of situations in which most people behave (e.g.) deplorably, so that these situations form no open list. I see, however, no evidence for this possibility.

character status: an indeterminate person is (e.g.) definitely not good.[81] So the claim that a person is indeterminate is in a way *dis*analogous to the claim that it's indeterminate whether Jane Eyre has any siblings, understood as the claim that *Jane Eyre* is silent on this question.[82]

Having thus clarified the concept of indeterminacy, I give now an argument for the claim that (Q2) fragmentation entails indeterminacy. For any (actual or hypothetical) person p:

(Q5) If p behaves deplorably in many situations, then p is not good.

(Q6) If p behaves admirably in many situations, then p is not bad.

(Q7) If p behaves deplorably in many and admirably in many other situations, then p is not intermediate (between good and bad).

Thus: (Q2) If p behaves deplorably in many and admirably in many other situations, then p is neither good nor bad nor intermediate.

(By "behaves" I mean "does or would behave" and by "many situations" I mean "an open list of actual or counterfactual situations".) The argument is clearly valid,[83] but

---

[81] An analogy may clarify my contrast between *undefined* and *silent*. Suppose I say: "consider a function f which to every nonzero rational number x assigns its inverse, $1/x$". Strictly speaking, my utterance is *silent* on what (if any) value f assigns to zero. But if it is implicitly understood that f assigns *no* value to zero, then f is *undefined* for zero; my utterance is then not silent, because it entails that f does *not* assign to zero the value (e.g.) 523.

[82] The claim that a person is indeterminate is *not* the claim that there is widespread disagreement on the question of what (if any) is the person's character status.

[83] Actually Q2 is *equivalent* to (rather than just *following* from) the conjunction of Q5, Q6, and Q7. Here is why. Q2 is clearly equivalent to the conjunction of Q7 with Q5′ and Q6′:

(Q5′) If p is fragmented, then p is not good.

(Q6′) If p is fragmented, then p is not bad.

But Q5′ entails Q5. In fact, suppose Q5′ is true and consider a person p_1 who behaves deplorably in an open list of situations. Take a person p_2 who (1) behaves exactly like p_1 (and thus deplorably) in an open list of situations included in the list of situations in which p_1 behaves deplorably, and (2) behaves admirably in all other (including an open list of) situations. Then p_2 is fragmented and is thus (by Q5′) not good. But p_1 never behaves better than p_2 and is thus not good either. Similarly, Q6′ entails Q6.

is it sound? One's answer will depend on which conception of character evaluations one adopts.

3.2. Three kinds of conceptions of character evaluations

(1) Conceptions of character evaluations according to which the argument for Q2 is sound may be called *consistency conceptions*. The label is apt because Q5 (similarly for Q6 and Q7) asserts a form of behavioral consistency: by contraposition, Q5 says that good people behave deplorably in at most *few* (i.e., a closed list[84] of) situations. Equivalently, Q5 precludes a form of *compensation*: it says that a person who behaves deplorably in many situations is not good, *regardless* of how admirably the person might *also* behave. To say that such a person is not good is not to say that she is bad: given also Q6, she may be neither good nor bad. So consistency conceptions strike a balance between two extreme positions on compensation: a "hard" line according to which a person who behaves deplorably in many situations is *bad* (*no* compensation is possible), and a "soft" line according to which such a person can even be *good* (*full* compensation is possible).[85] (2) Conceptions of character evaluations according to which the hard line is true may be called *impurity conceptions*: an open list of "impurities" (instances of deplorable behavior) guarantees badness. In contrast to consistency conceptions, which are *symmetric* in the sense of requiring consistency both of good and of bad people, impurity conceptions are *asymmetric*: they require consistency of good but not of bad people. (3) Finally, according to what may be called *averaging conceptions*, character evaluations function much like grade point averages: a student who gets many C's and many A's can still be a

---

[84] I am using 'few' in a special sense: a *closed list* of situations can consist of a *large number* of similar (and thus not multifarious) situations.

[85] More rigorously, the soft line is just the negation of Q5 and the hard line is: (H5) If p behaves deplorably in many situations, then p is bad. H5 is compatible with Q5 (in fact entails Q5) but is incompatible with Q6.

good or bad student if the A's far outweigh the C's or vice versa, and similarly a person who behaves deplorably in many and admirably in many other situations can still be good or bad. So averaging conceptions are symmetric (they require consistency neither of good nor of bad people)[86] and adopt a soft line on compensation. Table 2.1 summarizes the main characteristics of these three kinds of conceptions.[87]

	Q6 true (consistency required of bad people)	*Q6 false* (no consistency required of bad people)
Q5 true (consistency required of good people)	*Consistency conceptions* (symmetric; middle line on compensation)	*Impurity conceptions* (asymmetric; hard line on compensation)
Q5 false (no consistency required of good people)	____	*Averaging conceptions* (symmetric; soft line on compensation)

Table 2.1. Consistency, impurity, and averaging conceptions of character evaluations

According to impurity conceptions, fragmented people are bad rather than indeterminate; according to averaging conceptions, fragmented people may be good, intermediate, or bad, but are not indeterminate; it's only according to consistency conceptions that fragmented people are indeterminate. To defend consistency conceptions I will successively defend Q5 (§3.3), Q6 (§3.4), and Q7 (§3.5). My defense of Q5 will arguably

---

[86] Averaging conceptions can be asymmetric in the sense of giving greater weight to deplorable than to equally extreme admirable behavior.

[87] Further kinds of conceptions could be placed in the table but are not listed. For example, the cell of the table that corresponds to consistency conceptions (according to which Q7 is true) also corresponds to conceptions according to which Q5 and Q6 are true but Q7 is false.

make averaging conceptions implausible, but my defense of Q6 will be more tentative: impurity conceptions are not so implausible, but I will argue that still they are less plausible than consistency conceptions.

3.3. Prevalent deplorable behavior precludes goodness (Q5)

Let us first make clear what Q5 does and does not say. As we saw, there are two equivalent ways of regarding Q5: as asserting a specific form of behavioral *consistency*, and as precluding a specific form of *compensation*. (1) Although Q5 asserts the form of consistency according to which good people behave deplorably in at most *few* situations, Q5 does not say that good people *never* behave deplorably: Q5 does not assert *perfect* consistency.[88] Nor does Q5 say that good people usually behave in the exact *same* way (cf. Doris 1996: 60-1): there are many ways of behaving nondeplorably. Q5 does not even say that good people usually behave in various *admirable* ways: some good people may usually behave *neutrally*, neither deplorably nor admirably. (2) Although Q5 precludes *full* compensation of *deplorable* behavior in *many* situations, Q5 is compatible with the following four forms of compensation. (a) Compensation of *peccadilloes*. (b) Compensation of deplorable behavior in *few* situations. (c) *Partial* compensation (i.e., ascribing lack of badness rather than ascribing goodness) of deplorable behavior even in

---

[88] Cf. Mandelbaum (1955: 172): "A virtuous man need not be wholly without vices, nor an evil man without virtues." Four forms of g-consistency ('g' for 'good') can be distinguished:

G-consistency	Strong	Weak
Perfect	Always behaving admirably (i.e., never behaving non-admirably)	Never behaving deplorably
Moderate	Behaving nonadmirably (i.e., either deplorably or neither deplorably nor admirably) in at most few situations	Behaving deplorably in at most few situations

The form of consistency asserted by Q5 is weak moderate g-consistency, which is weaker than all three other forms of consistency. One can also distinguish four analogous forms of b-consistency ('b' for 'bad') by replacing "admirably" with "deplorably" and vice versa, and various forms of m-consistency ('m' for 'intermediate') by conjoining or disjoining various forms of g- and b-consistency. The weakest of these forms of m-consistency is the disjunction of weak moderate g- and b-consistency, so the most acute corresponding form of *in*consistency is the negation of this disjunction; it can be seen that this most acute form of inconsistency corresponds to fragmentation.

many situations. (d) *Diachronic* compensation (i.e., moral transformation): Q5 allows that a criminal can become a saint. This is because the antecedent of Q5 refers to *current* behavior and thus fails to be satisfied by a saint who *used* to behave deplorably but no longer does.[89, 90]

Given the above clarifications, Q5 should look plausible: goodness of character may be compatible with a *small* number of *mild* moral transgressions, but seems incompatible with a *large* number of *severe* transgressions. Note that Aristotle asserts something like Q5: "the decent person will never willingly do base actions".[91] This claim, like Q5, is about a "decent" person (επιεικής), namely a (non-superlatively) good person (cf. Irwin 1985: 392) rather than a "moral exemplar" (cf. Blum 1994). Similar claims (though sometimes about moral exemplars) figure prominently in neo-Aristotelian ethical thought (Doris 1996: 57-60, 1998: 506, 511-2). It seems then that I have plenty of company in finding Q5 intuitively appealing. Besides philosophers, this company in-

---

[89] Q5 is also compatible with the *im*possibility, in some cases, of moral transformation (maybe some criminals can never deserve canonization, because some crimes are too heinous to be ever expiated). This is because Q5 gives only a *sufficient* condition for failing to be good, and is thus compatible with the possibility that some kinds of past behavior provide *another* sufficient condition.

[90] Q5 should also be distinguished from: (R5) For any constellation of admirable behavior, there is a constellation of deplorable behavior such that anyone who exhibits both is not good. It can be seen that Q5 entails R5 but not vice versa. Note finally that Q5 is compatible with: (S5) There is a situation and a way of behaving deplorably such that, if *p* behaves in the given way in the given situation, then *p* is not good. The idea is that *some* kinds of deplorable behavior may be so extreme that their occurrence in a *single* situation suffices to preclude goodness: an *open list* of situations is not needed. Related to S5 is Mandelbaum's (1955: 172) claim that an *engulfing*—"single, controlling, pervasive"—vice can ensure badness. Given Q6, I reject Mandelbaum's claim (cf. Skowronski & Carlston 1992: 441-2), but I accept that an engulfing vice can preclude goodness (as opposed to ensuring badness).

[91] *Nicomachean Ethics* 1128b28-9; cf. 1100b35. This is a claim of *perfect* consistency, given Aristotle's use of the word 'never' ('ουδέποτε'; contrast 1100b19). Unlike this claim, Q5 is not about "willing" behavior: deplorable behavior can be willing (cf. the "sadistic" guards in Zimbardo's experiment) or unwilling (cf. the reluctantly obedient participants in Milgram's experiment). A related claim of Aristotle's is his (controversial) "reciprocity of the virtues" thesis (1144b31-1145a2): "you have one of the virtues of character if and only if you have them all" (Irwin 1988: 61; cf. Badhwar 1996; Doris 1996: 61-6, 1998: 521 n. 11; Flanagan 1991: 261-5, 282-3).

cludes ordinary people; this is suggested by certain psychological studies which I examine next.

3.3.1. Empirical evidence for Q5

Riskey and Birnbaum (1974) gave to each of 50 undergraduate students a booklet containing descriptions of 35 sets of actions, each set consisting of 2-11 actions. Each student was instructed to "read each set of actions and then judge how 'good' or 'bad' it would be to carry out *all* of the actions. ... In other words ... how morally 'commendable' or 'reprehensible' a person would be who carried out *all* of the actions." "Ratings were made on a 17-point scale ranging from -8 (very very bad) to +8 (very very good) in which zero was designated as neither good nor bad." A previous study (Birnbaum 1973) indicated that some of the actions described in the booklet were deplorable (e.g., secretly spiking a party's potato chips with a dangerous drug) while others were admirable (e.g., rescuing a family from a burning house). It was found, for example (see the lower right corner of Table 2.2), that a set consisting of two deplorable and nine (equally extreme) admirable actions (2D9A) got a mean rating of about -2.4 by the 50 students. The results reported in Table 2.2 support Q5 because, as the number of admirable actions which are added to two deplorable ones increases (up to nine), the mean ratings seem to approach a limit which is below -2; this suggests that a person who behaves deplorably is not evaluated as good, regardless of how admirably she also behaves. Similar results were obtained by Reeder and Coovert (1986); also, for evaluations of honesty rather than moral character, by Skowronski and Carlston (1992).

	0A	1A	3A	5A	7A	9A
2D	-7.3	-5.2	-3.6	-2.9	-2.5	-2.4

Table 2.2. Results from Riskey and Birnbaum (1974)

A related study was carried out by myself (Vranas 2000). I presented introductory psychology students with descriptions of hypothetical persons who perform various numbers of deplorable and admirable actions, and I asked the students to evaluate the persons in terms of moral character as good, intermediate, or bad.[92] I found (see the following table) that a person who performs a large number of deplorable actions (I gave three representative examples of such actions: 3D) and a much larger number of admirable actions (I gave nine examples: 9A) was evaluated as good by only 2 (5.3%) of 38 students. In other words, I found almost unanimous agreement that a person who performs many deplorable and many more admirable actions is not good.

% good	0A	3A	6A	9A
3D	0.0%	0.0%	2.8%	5.3%

Table 2.3. Results from Vranas (2000)

One might object that the above results are compatible with the possibility that a person who performs many deplorable and *enormously* more admirable actions would be evaluated as good by most people. I examined this possibility by means of another questionnaire item:

> Suppose Zed has performed (and is *still* performing) a large number of *very* bad actions (with corresponding—i.e., bad—*motivation*). Is there anything you could learn about Zed that would make you conclude that Zed is nevertheless a good person?

---

[92] Each student had also the options of responding that (a) the person's character could not be classified (was neither good nor bad nor intermediate) or (b) the description provided insufficient information to judge the person. I took care to specify that for good actions good motives and for bad actions bad motives were to be assumed. (Riskey and Birnbaum (1974) did not make explicit any assumption about motives, but arguably the context of their study conversationally implicated that the described actions were intentional, adequately informed, and so on.)

Almost 65% (i.e., 20) of 31 students answered "no", but agreement with Q5 is probably more widespread than this percentage suggests. Only one of the 11 students who answered "yes" gave as her reason that Zed might perform good actions in addition to bad ones; 7 of the remaining 10 students gave as their reason either that Zed's motives might be good or that Zed's character might change—so they misunderstood the question and they did not clearly disagree with Q5. On the other hand, it seems that only 3 of the 20 students who answered "no" misunderstood the question; most of the remaining 17 students said things like "you can't be good and perform so many bad actions".[93]

Some people might take issue with my use of psychological studies to support Q5: how can *empirical* results be relevant to a *non-empirical* claim like Q5?[94] To see how they can, take an example. Suppose a Gallup poll would indicate that only a tiny minority of people disagree with the assertion "all (actual or possible) midwives are female"; the overwhelming majority agree because they use (and they think that almost everyone uses) the word 'midwife' to refer only to women.[95] Then the claim that all (actual or possible) midwives are female would be probably true. In this example the empirical evidence *strongly* supports the non-empirical claim, but one might object that in other examples the support is much weaker: backward causation may well be conceptually possible even if a Gallup poll indicates that the overwhelming majority of peo-

---

[93] In another study (Reeder, Henderson, & Sullivan 1982) 20 college students were asked: "In general, how often does a very moral person act very immoral?" The mean response was about 2.4 on a scale ranging from "Never" (1) to "Very often" (7). (I performed a replication, with similar results: see footnote 106.) Twenty other students were asked: "If a large reward were available for doing so, how likely is it that a person who is very moral would try to act very immoral?" The mean response was slightly above 2 on a scale ranging from "Not very likely" (1) to "Very likely" (7). These results support Q5 because they suggest that a good person is expected to behave immorally only rarely, even when tempted to do so by a large reward (cf. Reeder & Brewer 1979: 68; Reeder & Spores 1983: 744).

[94] The objection assumes that there is a distinction between empirical and non-empirical claims and that Q5 is non-empirical (presumably because Q5 is about all *possible* persons). For the sake of argument I grant these assumptions.

[95] And because, though they know that dictionaries say otherwise, they think that 'midwife' *should not* refer to men.

ple think otherwise. Now if the people who think otherwise do so because of faulty reasoning (e.g., because they mistakenly believe that backward causation involves changing the past), then I agree. But if the empirical evidence indicates instead (analogously to the midwife example) that the people who think otherwise do so because they use the word 'cause' so that no (actual or possible) effect ever precedes any of its causes, then I disagree: such empirical evidence does strongly support the conceptual impossibility of backward causation.[96] I submit that in the case of Q5 we are closer to the latter kind of explanation: we have evidence on how people *use* the concept of goodness of character (the participants did evaluate hypothetical persons), and also on *why* people agree with Q5 (see the end of the last paragraph). Moreover, unlike the backward causation issue, Q5 is not overly complex or confusing: it's a universally quantified claim, like the claim that all midwives are female. It's true that subtleties are involved in judging particular actions as deplorable or admirable and in integrating information from various actions to evaluate a person's character; so the *possibility* exists that people's responses to the questionnaires rest on faulty reasoning. But I readily grant that the empirical results are not *conclusive* evidence for Q5; I claim rather that the results shift the burden of proof to *opponents* of Q5. I will examine now whether such opponents can shoulder this burden: I will address five objections to Q5.

3.3.2. Five objections to Q5

Objection 1: Good Motives. Recall (from §2.1) that whether an action is deplorable depends in general not only on the agent's motives but also on whether the action is wrong (in the sense of violating the agent's duty). But then one might object to Q5 that a

---

[96] Of course it's possible that people's responses would change if backward time travel became commonplace, but in such a case the concept of causation would *change* (compare the way in which the widespread acceptance of Special Relativity Theory changed the concept of time). See §3.6 for further discussion.

person's goodness of character should depend *only* on the person's motives; for example, Mandelbaum understands character as "the relatively persistent forms which a person's motivation takes" (1955: 141; cf. Brandt 1970/1992b; Doris 1998: 509-10) and states that "we frequently hold a conscientious person to be virtuous even though we deprecate the moral choices which he makes" (1955: 170). I agree with the second statement if the choices are only *mildly* immoral (hence not deplorable and not threatening Q5), but I disagree with the claim (which would falsify Q5) that prevalent *deplorable* behavior which is impeccably motivated is compatible with goodness. For the purposes of the latter claim, impeccable motivation cannot consist merely in a *de dicto* desire to do the right thing, because such a desire can coexist with horrifyingly immoral *de re* desires which preclude goodness: an anti-Semite may sincerely believe that exterminating Jews is morally right. But if impeccable motivation includes consistently moral *de re* desires, how can it correspond to deplorable behavior? Maybe through weakness of will:[97] isn't a person good if she consistently has moral *de re* desires but behaves deplorably because she is weak-willed? No: maybe such a person is not bad, but she is not good either. Consider: I don't want to beat my children but I keep losing my temper. I wanted to call the police when I witnessed a rape but I didn't bring myself to do it. I want to stop and help you but I'm overcome by my haste to go home and check my email. And so on. Then I'm not good despite my impeccable motivation. (Maybe my motivation is not "impeccable" because it lacks sufficient strength, but if impeccable motivation is understood as sufficiently strong then it cannot correspond to deplorable behavior and the current objection to Q5 does not even get off the ground.)

---

[97] One might think that another possibility is through misinformation: I may beat my children because I believe it's good for them. But then I do have an immoral *de re* desire, namely to beat my children. What if my misinformation is nonculpable, for example because I was given a drug that made me believe that beating one's children is good for them? I still have the immoral *de re* desire to beat my children; moreover, my behavior may be adequately excused and thus not deplorable (even if it is still wrong).

Objection 2: Extreme Behavior. One might grant that *extremely* deplorable behavior, like that of a serial killer, can be compensated for by no amount of admirable behavior, but might object that compensation is possible in less extreme cases: what about a person who regularly crushes ants just for fun but is otherwise a model citizen? I agree that such a person might still be good, but I think this will not do as a counterexample to Q5: I think that the habit of crushing ants just for fun is not deplorable (it's blameworthy but not *seriously* so), and that this habit may amount to behavior in a *single* recurrent situation rather than an *open list* of situations. Some people may not be convinced because they take crushing ants very seriously. But then we disagree about the *antecedent* of Q5, about which actions count as deplorable. I can live with such disagreement: in §2.2 I argued that the actions which are relevant to my purposes (e.g., nonsuspicious obedience in Milgram's experiment) are indeed deplorable. In response one might modify the example: what about a person who, in addition to crushing ants, also regularly kills squirrels and sets cats on fire just for fun? I agree that these new habits are deplorable, but I wouldn't call such a person good. But now one might complain that my strategy makes Q5 unfalsifiable: for any putative counterexample, I can maintain either that the behavior in question is not deplorable or that the person in question is not good. I hope indeed I can maintain so, but by appealing to claims *different* from Q5. For example, claims about deplorable behavior: I'm not defining deplorable behavior as behavior which precludes goodness. So it *is* in principle possible to find convincing counterexamples to Q5. I just haven't found any.

Objection 3: Extreme Situations. According to Q5, deplorable behavior in *any* open list of situations precludes goodness. One might object, however, that deplorable behavior in *extreme* situations is irrelevant to goodness: the fact that you would betray your country if you were tortured does not count against your being good. But what if under torture you would *gladly* betray your country? Not *every* behavior under torture is irrelevant to goodness. I claim that if wrong behavior under torture is irrelevant to good-

ness then it is also irrelevant to Q5. This is because wrong behavior under torture is irrelevant to goodness only if under torture the behavior is adequately *excused*;[98] but if it is, then it is not blameworthy, let alone deplorable (although it is by assumption wrong), so the antecedent of Q5 does not apply. The extremity of the situation is a red herring: what matters is the presence of an adequate excuse. Extremity need not provide an excuse, and in the absence of an excuse seriously wrong behavior in an extreme situation is relevant to goodness: the fact that in a fire you would inexcusably let your children perish does count against your being good. Similar remarks apply to *change-inducing* situations. If the fact that you would betray your country after being brainwashed does not count against your being good, this is not just because brainwashing would change your character: it's because the change would be *excusable*.[99] Deplorable behavior due to an inexcusable change is relevant to goodness: the fact that if you were to meet a certain person you would inexcusably become so infatuated with her that you would abandon your spouse and children does count against your being good.[100]

Objection 4: Unchosen Situations. How can the fact that you would behave deplorably in an open list of situations prevent you from being good if you manage to *avoid* these situations? Here is how. Consider a person p_1 who never goes to bars because he knows that if he did he would get into fights and would start shooting people. Suppose

---

[98] This claim is compatible with Brandt's (1969/1992a: 229; cf. 1970/1992b: 263) claim that wrong behavior is excused if it does not manifest some defect of character: I am not saying that it's features of situations regardless of considerations of character which provide excuses.

[99] The change might be excusable even if you were *voluntarily* brainwashed; then the fact that you would be brainwashed does count against your being good but the fact that you would betray if you were brainwashed still does not.

[100] A situation s can also be "change-inducing" in the sense that, if you behaved deplorably *once* in s, you would *not again* behave deplorably in (any situation qualitatively identical to) s. Does deplorable behavior in an open list of *such* situations preclude goodness? Yes: after any such change an open list of situations would remain in which you would (once) behave deplorably. But what about deplorable behavior in an open list of situations such that, if you behaved deplorably once in *one* of them, you would never again behave deplorably in *any* situation? *After* such a massive change you need not fail to be good, but I think (just as Q5 implies) that *before* the change you are *not* good.

also that p_1 studiously avoids being alone with little girls, including his own daughter, because he knows he would not resist the urge to molest them. And so on. Then p_1 is not good, even if he never *actually* behaves deplorably (and even if he often behaves admirably). In response one might claim that a person p_2 who has (e.g.) the urge to molest his daughter but would always successfully resist this urge need not fail to be good—and is even ceteris paribus *better* than a person p_3 who has no such urge.[101] I reply that my claim that p_1 is not good does not contradict the claim that p_2 may be good (and even better than p_3): although p_1 (like p_2) does not *actually* molest his daughter, by assumption p_1 (unlike p_2) *would not* successfully resist the urge to molest his daughter if he were alone with her. I can also grant that p_1 is ceteris paribus better than a person p_4 who would behave deplorably in the same situations as p_1 but does not avoid these situations (even if, as luck would have it, p_4 never finds himself in any of these situations); so I can accept an idea which I take to underlie objection 4, namely that a disposition to choose the right situations matters for goodness. But the objection neglects the fact that deplorable behavior in unchosen situations *also* matters.

Objection 5: Counterfactual Behavior and Moral Luck. It is a consequence of Q5 that even *counterfactual* prevalent deplorable behavior precludes goodness, and the example (of person p_1) I gave in response to objection 4 suggests that this consequence is true. One might argue, however, that this consequence is incompatible with something that Nagel says in his discussion of "moral luck": "We judge people for what they actually do or fail to do, not just for what they would have done if circumstances had been different" (1976/1979: 34).[102] I reply that there is no incompatibility because (as the

---

[101] This is so on a "battle citation model" of goodness of character: "an agent is creditable for performing a right act if and only if a morally good desire won a hard battle in the war against temptation" (H. Smith 1991: 281-2). It is a matter of debate whether Kant (*Groundwork* 398) adopts such a model (cf. Benson 1987; Henson 1979; Herman 1981).

[102] The word 'just' should be dropped from the quotation if Nagel's point is that counterfactual behavior is *irrelevant*; otherwise Nagel is saying that actual behavior is *also* relevant, and I need not disagree. Note

context makes clear) Nagel understands the 'judgments' in question as ascriptions of responsibility, not as character evaluations: his claim that "[a] person can be morally responsible only for what he does" (1976/1979: 34) is compatible with my claim that a person can fail to be good because of what he *would* do.[103] But even if Nagel is not *talking* about assessments of character, doesn't his point also *apply* to such assessments? No: I can grant that the point applies to "assessing an agent's moral worth for his performance of a particular act", but this "involves a very different judgment from assessing his overall moral virtue [i.e., goodness of character]" (H. Smith 1991: 289; cf. Herman 1981: 368-9). Counterfactual behavior can be decisive for the latter kind of assessment even if it is irrelevant to the former; as an analogy, you are not brave if in every dangerous situation *but one* you would behave as a coward, but you may still deserve a medal for having behaved as a hero in the *only* dangerous situation you have ever faced. I am not denying that *some* concepts are applied exclusively on the basis of actual behavior: you are a murderer exactly if you have murdered (Sabini & Silver 1982: 146). I am rather saying that being a good person is in a way more like being brave than like being a murderer: whether you are good depends in general not only on how you actually behave but

---

that "what [you] would have done if circumstances had been different" is irrelevant to Q5 if your character would have been different if circumstances had (e.g., if you had been raised in Nazi Germany): only what you would do given your *actual*, present character can be relevant to Q5.

[103] What Nagel (1976/1979: 28) calls "luck in one's circumstances" and "luck in the way one's actions and projects turn out" are irrelevant to character evaluations: the morally unlucky driver who fails to have his brakes checked and accidentally kills a child is ceteris paribus just as good or bad as the morally lucky driver who also fails to have his brakes checked but kills no child (even if only the former driver deserves blame and punishment). Consider also a real-life example. On a certain night James Russell Odom and James Clayton Lawson Jr., both former convicted rapists, "go out looking for a victim … They stop at a 7-Eleven on U.S. Highway 1 and spot a young woman they like working behind the counter. But too many people are around, so they leave … The next night … they drive back to the 7-Eleven. … This time, they're the only ones in the store, so they abduct the young female store clerk" (Douglas & Olshaker 1995: 174). It's reasonable to suppose that Odom and Lawson would have abducted the woman on the first night had there been no other customers in the store. If so, does the counterfactual abduction on the first night carry any less weight than the actual abduction on the second night in our assessment of Odon and Lawson's characters at the time of the abductions? (They are of course responsible only for the actual abduction. Note also that what Nagel calls "constitutive luck" can be relevant to the character one has and thus also to character evaluations.)

also on how you would behave in "morally dangerous" situations like temptations or provocations. (A reluctance to regard some counterfactuals as relevant to goodness may arise from uncertainty about their truth: how can I *know* that you would betray me *if* you were offered a bribe? Such uncertainty is clearly irrelevant to Q5, which is not an *epistemic* claim.[104])

This completes for the moment my defense of Q5. (See §5.2 for further discussion.) Given that averaging conceptions of character evaluations contradict Q5 (§3.2), my defense of Q5 makes such conceptions implausible. I turn now to Q6.

3.4. Does prevalent admirable behavior preclude badness (Q6)?

3.4.1. An argument against Q6: negativity effects

The symmetry between Q5 and Q6 might raise eyebrows. An extensive psychological literature documents the existence of asymmetries known as *negativity effects*: "impressions of character are more strongly influenced by negative than by positive information" (Richey & Dwyer 1970: 77). For example, in a study very similar to the one by Riskey and Birnbaum (1974) that I described in §3.3.1, Birnbaum (1973) found that a set consisting of two admirable and two (equally extreme) deplorable actions got a mean rating of slightly above 3 ("bad"), definitely below the midpoint (5, "neutral") of a scale ranging from 1 ("very very bad") to 9 ("very very good").[105] Such results, however, are compatible with Q6, which says that admirable behavior in an *open list* of situations precludes badness: *two* situations don't make up an open list, so it is compatible with Q6 that a person who behaves admirably twice is bad. Stronger evidence against Q6 is pro-

---

[104] One might argue that there is no fact of the matter about (e.g.) whether you would betray me. I need not take a stand: only counterfactuals about the truth of which there *is* a fact of the matter are relevant to Q5.

[105] Cf. Birnbaum 1972; Cusumano & Richey 1970; Kanouse & Hanson 1972; Lewicka, Czapinski, & Peeters 1992; Reeder & Coovert 1986; Richey, Bono, Lewis, & Richey 1982; Richey, Koenigs, Richey, & Fortin 1975; Richey, McClelland, & Shimkunas 1967; Richey, Richey, & Thieman 1972; Skowronski & Carlston 1987, 1989, 1992.

vided by a study carried out by myself (§3.3.1). I found (Table 2.4) that a person who performs a *large* number of admirable actions (I gave three representative examples of such actions: 3A) and a much larger number of deplorable actions (I gave nine examples: 9D) was evaluated as bad by 24 (63.2%) of 38 students.[106]

% bad	0D	3D	6D	9D
3A	0.0%	44.7%	51.4%	63.2%

Table 2.4. Further results from Vranas (2000)

Moreover, in my study I used the following questionnaire item:

> Suppose Zed has performed (and is *still* performing) a large number of *extremely* good actions (with appropriate—i.e., good—*motivation*). Is there anything you could learn about Zed that would make you conclude that Zed is nevertheless a bad person?

Almost 84% (i.e., 31) of 37 students answered "yes", and almost 84% (i.e., 26) of those who answered "yes" gave as their reason that Zed might perform bad actions in addition to good ones. But then it seems that Q6 is false: by carrying out the above study I have dug my own grave.

3.4.2. An argument for Q6: incommensurability

Things are not so simple, however, because there is also an argument in support of Q6:

---

[106] Another negativity effect is that, on average, bad people are expected to behave admirably more frequently than good people are expected to behave deplorably. In my study, e.g., to the questions "How often does a bad person perform *extremely* good actions?" and "How often does a good person perform *extremely* bad actions?" the mean responses of (the same) 32 students were 4.0 ("somewhat rarely") and 2.6 (between "almost never" and "rarely") respectively. Cf. Gidron, Koehler, & Tversky 1993; Rothbart & Park 1986; and the references in footnote 93.

(Q8) Every bad person is *worse* than every intermediate person.

(Q9) No fragmented person is worse than every intermediate person. (I.e.: for any fragmented person *f* there is an intermediate person *m* such that *f* is not worse than *m*.)

Thus: (Q6′) No fragmented person is bad. (Q6′ entails Q6: footnote 83.)

The argument is clearly valid. Q8 follows from the *meaning* of 'intermediate': *between* good and bad. Q9 follows from Q9′:

(Q9′) Every fragmented person is *incommensurable*—i.e., neither better nor worse nor equally good (or bad)—with every *paradigmatically intermediate* person (who *never* behaves deplorably or admirably).[107]

To examine Q9′ I used the following questionnaire item:

> Consider two people. Person 1 has performed a large number of *very* good actions and a large number of *very* bad actions; *all* of Person 2's actions have been intermediate, between good and bad. Which person has a better character?

Fifteen (48%) of 31 students said that the two persons "cannot be compared", 14 (45%) students said that Person 2 is better, and 2 (7%) students said that the two persons are "equally good (or bad)". It seems then that there is substantial disagreement concerning Q9′: a large percentage of people accept it, but another, about equally large percentage reject it. Are there any *theoretical* considerations for or against Q9′?

In support of Q9′ I can adduce a general argument for incommensurability: if an item is *much* better than another with respect to *many* components of a given value and is *much* worse with respect to *many* other components, then the two items are incommensurable with respect to the given value. Take, for example, two students: Student 1

---

[107] Q9′ is stronger than Q9, so why commit myself to Q9′? Because one can show (see Theorem 2.3 in §5.1) that *every indeterminate* (hence, if fragmentation entails indeterminacy, every fragmented) *person is incommensurable with every intermediate person.*

gets an A+ in a large number of (unrelated) courses and a C- in an equally large number of other courses (and gets a B in every remaining course), whereas Student 2 gets a B in every course. Who is a better student? Given that each of the two students is in *many* respects *much* better than the other, it does not seem right to say that either of them is overall a better student than the other. But saying that they are equally good does not seem right either: they are too different to be compared.[108] It's true that one can define a *scale* with respect to which they are equal. For example, they have the same GPA (exactly B). Still, they are not equally *good students*: even if the GPA scale *normally* captures the evaluative concept of a good student, in the *particular* case of Student 1 it does not.[109] A fragmented student (or person) is in a way like an *idiot savant* who is an idiot in many respects but a genius in many others: such a person cannot be compared with average ordinary people in terms of overall intelligence.[110]

In response one might grant that the above two students (each of whom has a GPA of B) are incommensurable but might invite me to consider (cf. Chang 1997: 17) Student 3, who gets an A+ in a large number of courses and an *F* (rather than C-) in a much *larger* number of other courses (and thus has a GPA much lower than B). Isn't Student 3 worse than Student 2? The point might seem even clearer for character evaluations: Hitler performed so many and so extreme deplorable actions that even if he also performed a smaller (but still large) number of less extreme admirable actions he

---

[108] There is another possibility: the two students are "on a par" (Chang 1997: 4-5, 25-7), "roughly comparable" (Parfit 1984: 431), "roughly equal" (Griffin 1986: 81). But this is implausible: each student is in many respects *much* better than the other. In any case, if every fragmented person is on a par with—hence not worse than—some intermediate person, Q9 still follows.

[109] So the case of the GPA scale does not show that my general argument for incommensurability (which refers to *values* rather than *scales*) is invalid. But if the argument is somehow invalid, I take it to be inductively strong: *most* evaluative concepts, like that of a good student, are not calibrated with respect to weird cases like that of Student 1.

[110] The analogy may be unfair: unlike evaluations of character, evaluations of ability exhibit *positivity* effects (cf. Martijn, Spears, Van der Pligt, & Jakobs 1992; Reeder & Brewer 1979; Skowronski & Carlston 1987, 1989, 1992; Trafimow 1997). Still, I think the analogy is useful.

was still bad—and worse than any paradigmatically intermediate person. In reply I am not going to deny that Hitler was bad: I don't know enough about his admirable actions or dispositions (if any). Consider instead Schitler, a hypothetical dictator who orchestrates the murder of several million people but also behaves admirably in an open list of situations. I am not saying that he is just nice to his friends and family: admirable behavior is by definition (§2.1) *highly* praiseworthy, like risking one's life to help an electrocuted stranger. Nor am I talking about a few isolated instances of admirable behavior: an open list comprises a *large* number of *multifarious* instances. But once these clarifications are really taken in, it becomes plausible to claim that a paradigmatically intermediate person, who *never* behaves admirably, is not better than Schitler. I am probably not alone in finding this claim plausible. In my study I asked participants to compare a paradigmatically intermediate person with a person who performs a large number of admirable and a *much larger* number of deplorable actions (I listed three and nine representative actions respectively). Ten (29%) of 35 students said that the two persons cannot be compared, 7 (20%) said that they are equally good (or bad), and 18 (51%) said that the latter person is worse. So almost half of my respondents did *not* evaluate the latter person as worse, even though he was described as having "killed ten people whose political views he disliked by sending mail bombs to them", "forced his daughter to become a prostitute", and so on.

3.4.3. The verdict on Q6 and a problem for impurity conceptions

We are left with a complicated state of affairs: there are considerations both for and against Q6. But if Q5 is true, why should Q6 be false? Deplorable and admirable actions are defined symmetrically (e.g., the former violate whereas the latter exceed one's duty), so I see no principled reason for adopting an asymmetric conception of character evaluations, one that treats deplorable and admirable actions differently. One might try, however, to justify such an asymmetry by arguing that it's (all-things-considered) im-

permissible to violate a (pro tanto) duty as a means to performing an admirable action; e.g., it's impermissible to kill someone so as to save someone else's life. It follows that avoiding violations of one's duties (and thus avoiding deplorable actions) takes precedence over performing admirable actions and thus should be given greater weight in character evaluations—or so the objection goes. I reply that it's not *always* impermissible to violate a duty as a means to performing an admirable action. It depends on how admirable the action is relative to the strength of the duty: it's impermissible to *kill* a firefighter so as to enter a burning building and rescue three *cats*, but it's permissible to *hit* a firefighter so as to rescue three *children*. The impermissibility, I submit, obtains only when avoiding a deplorable action takes precedence over performing a *less extreme* admirable one, and it is consistent with symmetry that the former should be given greater weight in character evaluations than the latter. One might respond that it's impermissible to *kill* a firefighter so as to rescue three *children*, although the admirable action of rescuing *three* people is more extreme than the deplorable action of killing *one* person. I reply that if the latter action is not adequately excused by being performed as a means to performing the former, then I deny that the former action is more extreme. But then aren't I committed to some kind of asymmetry? Maybe, but this would be an asymmetry between deplorable and admirable *actions* which count as equally extreme, and would not commit me to an asymmetry in *character evaluations* between *equally* extreme deplorable and admirable actions.[111] I conclude that this attempt to justify an asymmetric conception of character evaluations fails. Given also that I find the incommensurability argument for Q6 convincing, I accept Q6.

---

[111] Some of the data which undermine Q6 are based on evaluations of persons who perform deplorable and admirable actions which a group of students judged on average as about *equally* extreme. Judgments of extremity, however, are unreliable due to "context effects"; e.g., a deplorable action is judged on average as more extreme when presented in a context of *mildly* deplorable actions than when presented in a context of *highly* deplorable ones (Parducci 1968; cf. Marsh & Parducci 1978). The possibility of context effects, on the other hand, might weaken my support of Q5.

Given my acceptance of Q6, I would like to explain away the data which undermine Q6 as due to some sort of bias. I am not in a position to offer a well-supported debunking explanation of negativity effects, but it is possible that such effects are due to *affective* factors: even when a deplorable and an admirable action are evaluated as equally extreme, the deplorable action may have a greater emotional impact on the evaluator than the admirable one, with the result that a person who performs both actions is evaluated negatively rather than neutrally. Whether this explanation is adequate is an open question.[112] Now given that I am discarding without adequate justification the data which undermine Q6, why couldn't a proponent of averaging conceptions reject Q5 by discarding the data (§3.3.1) which support Q5? Because, as I will argue in a moment (§3.5), incommensurability considerations *support* Q5. In contrast to Q6, which is supported by some considerations but undermined by others, Q5 is supported by two kinds of considerations.

My case for Q6 being to a certain extent tentative, impurity conceptions are not utterly implausible. Those who find impurity conceptions attractive, however, should note that such conceptions still fall prey to my epistemic critique of appraisal *respect* (though not of *contempt*): on impurity conceptions fragmented people are bad rather than indeterminate, so if the rest of my argument goes through then evaluations of people as *good* (or intermediate) are epistemically unwarranted.[113]

---

[112] Cf. Peeters & Czapinski 1990; Sedikides & Skowronski 1993. In support of this explanation note that we may be reluctant to evaluate as bad those whose good sides are vivid to us. Cf. Ann Rule's (1989) biography of the serial killer Ted Bundy: Rule first met Bundy when they were both working at a crisis clinic.

[113] The conclusion that most people are bad may be unpalatable to proponents of impurity conceptions: in my study only one (3%) of 31 students said that most people are bad, whereas 12 (39%) said that most people are good, 15 (48%) said that most people are intermediate, and 3 (10%) said that most people are neither good nor bad nor intermediate.

3.5. Does fragmentation preclude being intermediate (Q7)?

Given Q5 and Q6, a fragmented person is neither good nor bad. But it doesn't yet follow that such a person is indeterminate: the person may be intermediate, between good and bad. Q7 excludes this possibility. One might think that Q7 follows from Q9′: if every fragmented person is incommensurable with every paradigmatically intermediate person, doesn't it follow that no fragmented person is intermediate? It doesn't: the possibility exists that some intermediate people are incommensurable with each other. I will argue, however, that the general argument for incommensurability which I used to support Q9′ also supports Q7. Here is my argument for Q7:

(Q10) Every intermediate person is worse than every good person.

(Q11) No fragmented person is worse than every good person. (I.e.: for any fragmented person *f* there is a good person *g* such that *f* is not worse than *g*.)

Thus: (Q7) No fragmented person is intermediate.

The argument is clearly valid. Q10 (like Q8) follows from the meaning of 'intermediate'. (In conjunction with Q10, Q9′ entails Q5; this is why I said in §3.4.3 that incommensurability considerations support Q5.) In support of Q11, take any fragmented person *f* and consider a good person *g* who (a) behaves admirably in every situation in which *f* behaves deplorably and (b) behaves neutrally (i.e., neither deplorably nor admirably) in every other situation. (Arguably *every* person who satisfies (a) and (b) is good, but I am using only the weaker claim that at least *one* such person is good.) Now recall (from §3.4.2) my general argument for incommensurability:

> If an item is *much* better than another with respect to *many* components of a given value and is *much* worse with respect to *many* other components, then the two items are incommensurable with respect to the given value.

This argument applies to the case at hand: *g* behaves much better than *f* in an open list of situations (in which *f* behaves deplorably and *g* admirably) and behaves much worse than *f* in another such open list (in which *f* behaves admirably and *g* neutrally). It follows that *f* is incommensurable with—hence not worse than—*g*, so that Q11 is true.

Besides being supported by the above formal argument, Q7 is intuitively plausible. Take an analogy. Being between hot and cold amounts to having a mild (intermediate) temperature. But a "fragmented" lake, which has very many hot and very many cold areas, does not have a mild temperature: it has no overall temperature. One might object that this analogy is inappropriate: character evaluations combine a multitude of factors, whereas temperature is in a sense a single factor. Take then another analogy: my attitude towards gun control combines a multitude of factors, namely various considerations for and against. If I believe that only a *single* kind of guns should be available to private citizens but there should be *no* restrictions on who can own guns of this kind (cf. LaFollette 2000: 263), then it's inaccurate to say that my attitude is between for and against: I have a complicated attitude, which cannot be properly placed on a for/against scale. One might object that it would not be strange for me to choose the midpoint of a scale when questioned about my attitude towards gun control. I have two replies. First, the midpoint can be ambiguous. Recall that in Riskey and Birnbaum's (1974) study "zero was designated as neither good nor bad"; thus zero could be understood as corresponding either to being between good and bad or to being indeterminate.[114] Second, we frequently choose midpoints not because they are appropriate, but because of situational pressures. Take a student paper which exhibits both outstanding

---

[114] A similar situation is typical: in the semantic differential (a series of bipolar adjective scales), "the most popular way of measuring attitudes in contemporary research", "[t]ypically the instructions tell the respondent to check the middle category if neither adjective describes the object better than the other *or if both are irrelevant* to it" (Himmelfarb 1993: 55-6, italics added).

originality and disheartening reasoning mistakes. If we *have* to give the paper a grade, then an intermediate grade may be the most reasonable option. But we may well feel that the paper is not properly characterized as being between good and bad. One might object that such a paper is still worse than a paper which is terrific and better than a paper which is terrible on *all* counts (including both originality and reasoning); similarly, every fragmented person is worse than every *perfectly good* person (who always behaves admirably) and better than every *perfectly bad* person (who always behaves deplorably). I reply that (as Q8 and Q10 say) to be intermediate a person must be between *every* good and bad person; being between every *perfectly* good and bad person does not guarantee being intermediate, because even some of those who are good or bad are between all of those who are perfectly good and all of those who are perfectly bad.

3.6. Do consistency conceptions correspond to considered *opinions?*

To support the claim that (Q2) fragmentation entails indeterminacy, I appealed to evaluations of hypothetical persons by participants in psychological studies. But would these participants persist in their evaluations if the prevalence of fragmentation were made salient to them? One might argue that they would not: if they were to realize that combining consistency conceptions with the results of certain experiments leads to the conclusion that most people are indeterminate, then they would disavow consistency conceptions. For example, they would evaluate certain persons who behave deplorably in many situations as intermediate, because they would take into account that prevalent deplorable behavior characterizes *most* people. So consistency conceptions are at most superficially appealing and do not correspond to our *considered* opinions—or so an objector might argue.

In reply I contest both the premise and the validity of the above argument. (1) The premise says that the participants would change their evaluations if the prevalence of

fragmentation were made salient to them;[115] but how do we know that they would? The experimental results which support the claim that most people are fragmented are surprising; so why shouldn't they have a surprising consequence, namely that most people are indeterminate? The unexpectedness of this consequence need not lead to a disavowal of consistency conceptions, given that it can be traced to the corresponding unexpectedness of the experimental results. (2) But even if I grant that the evaluations *would* change, it doesn't follow that Q2 is false. The possibility exists that making the prevalence of fragmentation salient to people would cause them to *give up* consistency conceptions and *adopt* (rather than *upholding*) a non-consistency conception; this would be a case of conceptual change. In such a case the *sentence* 'fragmentation entails indeterminacy', which currently expresses Q2, would express some *other* proposition because our use of expressions like 'goodness of character' (and thus 'indeterminacy') would have changed; but the falsity of *that* proposition does not entail the falsity of Q2. In response one might grant that a consistency conception *may* correspond to our considered opinions even if we would give it up, but might ask why we should believe that it *does* correspond to our considered opinions if we would give it up. I reply that I supported consistency conceptions not only by appealing to evaluations of hypothetical persons but also by means of a general argument for incommensurability. My appeal to psychological studies had the limited purpose of shifting the burden of proof to opponents of Q5; the thrust of my reasoning consisted in rebutting objections to Q5 and in defending the argument for incommensurbility.

This completes my defense of the claim that (Q2) fragmentation entails indeterminacy, and thus of the claim that (L0) *most* people are indeterminate.

---

[115] A variant of the above argument does not talk about change of evaluations but rather says that participants to whom the prevalence of fragmentation is made salient will produce evaluations discordant with consistency conceptions to start with. My reply applies to this variant as well.

4. The posterior probability of indeterminacy

4.1. The real *argument for the epistemic thesis*

If most people are indeterminate, does it follow that our everyday practice of evaluating people in terms of their character as good, bad, or intermediate is epistemically unwarranted? No: the possibility exists that we can *reliably distinguish* the minority of people who are good or bad from the majority who are indeterminate. (As an analogy, although most people don't have Ph.D.'s, we can reliably distinguish the minority of people who have Ph.D.'s from the majority who do not.) So I need to do something more if I am to conclude that character evaluations are epistemically unwarranted: I need to argue that we cannot reliably distinguish those who are indeterminate from those who are not. This is my task in the current section. More formally, what follows from the claim that most people are indeterminate is that (L1) our *prior* probability that a randomly selected person is indeterminate should be high.[116] In this section I argue that, even when our evidence about a person is taken into account, our *posterior* probability that the person is indeterminate should also almost always be high.

Contrary to what I advertised in §1, my argument for the conclusion that the posterior probability of indeterminacy is high will *not* go through an argument that the posterior probability of *indeterminacy* should not be much lower than the prior. I will argue instead that the posterior probability of *fragmentation* should not be much lower than the prior, so that the posterior probability of *fragmentation* should be high, and thus so should be the posterior probability of *indeterminacy*, given that fragmentation entails

---

[116] One might object that this inference derives an "ought" from an "is": even if (L0) most people are in fact indeterminate, why should someone unfamiliar with the experimental results I presented in §2 have a high prior probability that a randomly selected person is indeterminate? She should only if she justifiably *believes* L0, but the premise of my argument from L0 to L1 is that L0 is *true*, not that anyone believes it. I reply that L1 and the epistemic thesis should not be understood as making claims about what particular people (who may or may not be familiar with the experimental results) are justified in believing. They should rather be understood in analogy with the claim that, given the scientific evidence for evolution, we should not believe in creationism: particular people may not be familiar with the scientific evidence, but the claim is about the beliefs that the evidence objectively supports. (Cf. also the lottery analogy in §1.)

indeterminacy. So although I think that talking about the prior and posterior probabilities of indeterminacy was suited for expository purposes, my real argument for the epistemic thesis is not the one I presented in §1 but is rather the following:

(Q1) Most people are fragmented.

Thus (from Q1): (L1′) For any p, $P(Fp)$ should be high.

(P1′) For any p, if the Independence Condition holds, then $P(Fp|Ep)$ should *not* be appreciably lower than $P(Fp)$.

(P2) For almost any p, the Independence Condition holds.

Thus (from P1′ and P2): (L2′) For almost any p, $P(Fp|Ep)$ should *not* be appreciably lower than $P(Fp)$.

Thus (from L1′ and L2′):(C1′) For almost any p, $P(Fp|Ep)$ should be high.

(Q2) Fragmentation entails indeterminacy.

Thus (from C1′ and Q2):(C1) For almost any p, $P(Ip|Ep)$ should be high.

In §4.2 I defend P1′. In §4.3 I defend P2. In §4.4 I address an objection.

4.2. Posterior vs prior probability of fragmentation

Consider a person p and let the random variables B_s (s = 1, …, S) take the values (e.g.) -1, 0, and +1 depending on whether p (does or would) behave deplorably, neutrally, or admirably in situation s. Let the *Strict Independence Condition* be that B_1, …, B_S are (jointly) independent; informally, the probability that (e.g.) p would behave deplorably in Milgram's experiment given that p would behave *deplorably* in the electrocution experiment is the same as the probability that p would behave deplorably in Milgram's experiment given that p would behave *admirably* in the electrocution experiment. Let the *Symmetry Condition* be that B_1, …, B_S are identically distributed; e.g., p is equally likely to behave deplorably in Milgram's experiment and in the electrocution experiment. These two conditions are clearly false, but they will be relaxed later on. For the moment I want to make the preliminary points that, when both conditions hold, the claim that the

posterior probability of fragmentation is not much lower than the prior (a) is still not trivial but (b) is nevertheless true.

Let F be the proposition that p is fragmented, and let the proposition E describe our evidence about p; e.g., E may state that p behaved deplorably in situation s. One might think that, if the Strict Independence Condition holds, then trivially our posterior and prior probabilities $P(F|E)$ and $P(F)$ should be equal: if learning that p behaved deplorably in s should leave unaffected our probabilities that p would behave deplorably (or admirably) in any other situation, then how could it affect our probability that p is fragmented? This reasoning is fallacious because learning that p behaved deplorably in s may affect our probability that p is fragmented by increasing our estimate of the *number* of situations in which p would behave deplorably, even if it leaves unaffected our probability that in any *specific* situation p would behave deplorably. As an analogy, if I know that ten independent tosses of a fair coin took place, then learning that the coin came up heads in tosses 1 through 6 increases (from five to eight) my estimate of the number of tosses among the ten in which the coin came up heads although it leaves unaffected my probability that the coin came up heads in any particular toss from 7 to 10. So more work is needed to show that, if the Strict Independence Condition holds, our posterior probability that F is fragmented should not be much lower than the prior. This extra work is partly carried out by the following theorem.

Theorem 2.2. *Consider S independent and identically distributed random variables $B_1, \ldots, B_S$, each of which can take the values -1, 0, and +1 with probabilities p_D, p_N, and p_A respectively ($p_D+p_N+p_A=1$). Let F be the event that more than σ_D of these variables take the value -1 and more than σ_A take the value +1. Let D_s be the event that B_s takes the value -1. Then: $P(F|D_s)-P(F) = (1-p_D)P$(exactly σ_D of the remaining $S-1$ variables take the value -1 and more than σ_A of them take the value +1)$-p_AP$(exactly σ_A*

of the remaining S-1 variables take the value +1 and more than σ_D of them take the value -1).[117]

Given that σ_D and σ_A are very large (they correspond to *open lists* of situations), the probability that *exactly* σ_D or σ_A variables take a specific value is very small, so both terms of the difference which gives us $P(F|D_s)$-$P(F)$ are very small and so is $P(F|D_s)$-$P(F)$. Of course $P(F|E)$-$P(F)$ may still be large if our evidence E is not limited to the claim that (D_s) p behaved deplorably in a *single* situation;[118] but given that in real-life our evidence about people is almost never so extensive as to encompass a number of situations approaching an open list, Theorem 2.2 suggests that $P(F|E)$-$P(F)$ will almost always be small (if Strict Independence and Symmetry hold). One might object that we may have observed some of our intimates for decades. I reply that our observations typically consist of a small number of recurring situations; for example, typically one major common factor in most of our observations is our own presence. One might also object that we can construct an inductively strong argument from how someone has behaved in a number of situations to how she would behave in other situations. In reply in §4.3 I

---

[117] I will give a proof of the following simpler result (the proof of the theorem uses the same methods): *If Δ is the event that more than σ_D of the S variables take the value -1, then $P(\Delta|D_s)$-$P(\Delta)$ = $(1-p_D)P$(exactly σ_D of the remaining S-1 variables take the value -1).* Let Δ_n be the event that exactly n of the S variables take the value -1; with an obvious change in terminology, Δ_n is the event of n "successes" in S "trials" of a Bernoulli process with success probability p_D and failure probability $q_D = 1-p_D$. Clearly $\Delta = \cup_{n>\sigma D} \Delta_n$, so (1) $P(\Delta|D_s)$-$P(\Delta)$= $\Sigma_{n>\sigma D}$ $[P(\Delta_n|D_s)$-$P(\Delta_n)]$. Now $P(\Delta_n D_s)$ is the probability of n successes in S trials *and* a success at the s-th trial, so it is the probability of n-1 successes in S-1 trials *times* the probability of a success at the s-th trial. So (2) $P(\Delta_n|D_s)$= $P(\Delta_n D_s)/P(D_s)$= $P(n$-1 successes in S-1 trials). Now (3) $P(\Delta_n)$= $P(n$ successes in S trials) = $P(n$-1 successes in S-1 trials and a success at the remaining trial)+$P(n$ successes in S-1 trials and a failure at the remaining trial) = $P(n$-1 successes in S-1 trials)p_D+ $P(n$ successes in S-1 trials)q_D. From (2) and (3) we get: $P(\Delta_n|D_s)$-$P(\Delta_n)$ = $q_D[P(n$-1 successes in S-1 trials)-$P(n$ successes in S-1 trials)]. So $P(\Delta|D_s)$-$P(\Delta)$= $\Sigma_{n>\sigma D}$ $q_D[P(n$-1 successes in S-1 trials)-$P(n$ successes in S-1 trials)] = $q_D\{[P(\sigma_D$ successes in S-1 trials)-$P(\sigma_D$+1 successes in S-1 trials)]+ $[P(\sigma_D$+1 successes in S-1 trials)-$P(\sigma_D$+2 successes in S-1 trials)]+…+ $[P(S$-1 successes in S-1 trials)-$P(S$ successes in S-1 trials)]$\}$ = $q_D P(\sigma_D$ successes in S-1 trials). □

[118] Trivially, for example, in the extreme case in which E states that p behaved deplorably in more than σ_D and admirably in more than σ_A situations, E *entails* that p is fragmented; so $P(F|E)$ should be 1 even if $P(F)$ is not 1.

give evidence against the inductive strength of such arguments. But if I am right then why were my own arguments in §2 from the existence of three situations in which most people would behave deplorably (or admirably) to the existence of an open list of such situations inductively strong? Because my arguments were about how *most* people would behave, not about how any *particular* person would behave, in an open list of situations.

I conclude that, if Strict Independence and Symmetry hold, then the posterior probability of fragmentation should not be much lower than the prior. Now let me relax the Strict Independence and Symmetry Conditions. Relaxing Symmetry does not affect my argument: it becomes hard to state explicitly a theorem analogous to Theorem 2.2 (because the probabilities p_D, p_N, and p_A are replaced by situation-specific probabilities p_{Ds}, p_{Ns}, and p_{As}), but the essential point remains that $P(F|E)$-$P(F)$ is small if E describes behavior in a *small* number of situations (relative to σ_D and σ_A). To relax Strict Independence, replace it with the *(Approximate) Independence Condition*, which states that $B_1, \ldots, B_S$ should be considered *approximately* independent. I can offer no rigorous argument to exclude the possibility that $P(F|E)$-$P(F)$ becomes large when strict independence is replaced with approximate independence, but it would be strange if such a discontinuity existed. Perhaps more seriously, one might worry that replacing the Strict with the Approximate Independence Condition is small improvement: if the former is false, why should the latter be true? It is to this question that I now turn.

4.3. Evidence for the Independence Condition

An extensive literature in personality psychology suggests that the average correlation coefficients between people's behaviors in various situations are low. Here are four examples (cf. Mischel 1968). (1) A massive series of studies was carried out in the late 1920s by Hartshorne, May, and their collaborators (Hartshorne & May 1928; Hartshorne, May, & Maller 1929; Hartshorne, May, & Shuttleworth 1930). They subjected thousands of schoolchildren to a battery of tests of (e.g.) honesty and found correlations

of on average about .26 (Hartshorne & May 1928: 383) between scores on these tests. In other words, a child who (e.g.) behaved much more dishonestly than average on a cheating test typically did not behave much more dishonestly than average on a stealing test. (2) Newcomb (1929) measured behaviors presumably reflective of introversion-extroversion in 51 "problem boys" and found an average correlation coefficient of .14 among these measures. (3) Dudycha (1936) examined the punctuality of 307 college students in six situations (coming to class, coming to an appointment, coming for breakfast, coming to church, etc.). He found an average correlation coefficient of .19 among punctuality scores in these situations. (4) For a more recent example, Peake (1982) monitored 63 undergraduates at Carleton College over a ten week period for behaviors related to friendliness and conscientiousness. He found average correlation coefficients of .08 for conscientiousness and .05 for friendliness.

One might object that these results conflict with the commonsensical observation that we *can* reliably predict our friends' behavior. In reply distinguish (as social and personality psychologists do) the *cross-situational consistency* from the *temporal stability* of behavior. The above results do not deny that behavior is temporally stable, namely that people behave in more or less the same way when the *same* situation recurs. In fact, the average correlation coefficients between repeated measures of people's behavior in recurring situations are typically much higher than cross-situational consistency correlation coefficients. In everyday life we typically observe our friends in a limited number of recurring situations; this is why we can reliably predict their behavior. But in unusual situations our predictions would often go awry. This claim is also supported by another psychological literature which suggests that personality characteristics don't enable us to reliably predict who helps in bystander intervention experiments or who obeys in obedience experiments. Latané and Darley, for example, found that personality variables such as "[a]lienation, Machiavellianism, acceptance of social responsibility, need for approval, and authoritarianism did not predict the speed or likelihood of help" (1970b:

116). In a more recent review of the helping literature by Piliavin et al. (1981: chap. 8), only a couple of personality traits were found to consistently predict helping behavior. Similarly, in a comprehensive review of the obedience literature, Blass concluded that, although personality variables can predict obedience, "some of the findings are either contradictory or weak" (1991: 399; cf. Meeus & Raaijmakers 1995: 168; Modigliani & Rochat 1995: 121-2).

Another objection to my argument for the Independence Condition is that a low correlation, or even a *zero* correlation, does not amount to independence. This is because a correlation coefficient is a measure of the degree of *linear* dependence between two variables; the coefficient can be zero even when there is a perfect non-linear dependence. I reply that the Independence Condition, as stated in §4.2, does not say that people's behaviors in various situations *are* independent of each other; it says rather that they *should be considered* approximately independent. To see the difference, take an analogy. Consider a pseudo-random number generator: a device that uses a deterministic algorithm to generate numbers whose distribution looks random. If I don't know (and I have no way of finding out) the algorithm that the device uses, then I should consider the distribution as random, although the distribution is not *in fact* random. Similarly, I claim that *given the current state of the art* in prediction methods, the low correlations have the consequence that *we should consider* people's behaviors as approximately independent; I cannot rule out the possibility that future prediction methods will enable us to distinguish those who are fragmented from those who are not.

A third objection to my argument for the Independence Condition is that the correlation coefficients, low as they are, are *not sufficiently low* to justify considering people's behaviors in various situations as even approximately independent. Funder and Ozer, for example, argue that some typical correlation coefficients in personality psychology are of about the same magnitude as the correlation coefficients that correspond to some typical social psychological experiments. In Milgram's obedience experiments,

for example, the correlation between experimenter proximity and obedience was calculated by Funder and Ozer to be .36 (1983: 110). Three points in reply. First, I am not making a *comparative* claim about the relative importance of personality characteristics versus situations when I defend the unpredictability of behavior in unusual situations. Second, quoting Nisbett (1980), Funder and Ozer (1983) took correlation coefficients with values up to .40 to be typical of personality research; but as we saw above, the average correlations found by Hartshorne and May (1928), Newcomb (1929), Dudycha (1936), and Peake (1982) were much lower than .40. Third, and most important, Ross and Nisbett (1991: 109-115; see also Ross & Thomas 1986, 1987) argue extensively that average correlation coefficients of almost .2 result in only negligible improvements in the predictability of behavior.

4.4. Epstein's objection from aggregation

In a series of publications, Epstein (1977, 1979a, 1979b, 1980, 1983a, 1983b, 1984, 1986) has adduced certain considerations that might be taken to provide an objection to my argument for the Independence Condition. Epstein basically argues that, although the average correlation coefficient for B_s is indeed low, this is only to be expected given measurement errors; the average correlation coefficient for Y_σ will be high, where each Y_σ is an *aggregate* measure of behavior over a large number of situations (i.e., Y_σ is the average of B_s over all s in the set σ of situations).[119] Epstein's considerations can be divided into two parts: an intuitive *a priori* argument, and a more rigorous argument based on the Spearman-Brown formula. I will address these two arguments in turn; then I will examine empirical evidence relevant to aggregation.

---

[119] One could equivalently define Y_σ as a sum (rather than an average): the correlation coefficients would remain unchanged.

4.4.1. Epstein's intuitive argument

Epstein writes (1979: 1102):

> It is no more reasonable to assess the stability of nonintellective behavior by correlating single observations than it is to assess the stability of intellective behavior by correlating single items in an intelligence test. Thus, for statistical reasons alone, the low correlations cited as evidence against the existence of stable response dispositions can be discounted.[120]

To take another analogy, if we measure the *lengths* of all objects in a series by using an imprecise instrument, then any two series of measurements are expected to correlate only slightly; but if we take repeated measurements and average them, then any two series of average measurements are expected to correlate highly. In terms of the theory of measurement, we can say that $B_s = T_s + E_s$, where B_s is the measured score, T_s is the "true score", and E_s is the measurement error (e.g., Gulliksen 1950). The average Y_σ of B_s over all s in σ will be approximately equal to the average of T_s if (as is standardly assumed) the average of E_s is approximately zero. But the average of T_s is just t_p, the length of object p; so the correlation coefficient between Y_σ and $Y_{\sigma'}$ (for two different sets σ and σ' of measurement-situations) will be approximately equal to the correlation coefficient between t_p and t_p, namely 1.

The problem with the above argument is its reliance on the assumption that T_s is constant across s. This assumption seems reasonable when it's *length* that we are measuring (after all, if we take standard precautions, we don't expect the length of an object to change each time we make a measurement), but the corresponding assumption when we are measuring *behavior* results in a circular argument. Opponents of cross-situational consistency maintain that the true scores T_s and $T_{s'}$ are *not* equal because person p does

---

[120] Epstein is talking here about the *temporal stability*, not about the *cross-situational consistency* of behavior (see Epstein 1983a), but an analogous argument (as Epstein himself points out) might be given about cross-situational consistency.

not behave in the same way in situations s and s' (even when we abstract from questions of measurement error). So in the context of the cross-situational consistency debate Epstein is not entitled to assume that T_s and $T_{s'}$ are equal, and thus his argument fails.

4.4.2. Epstein's use of the Spearman-Brown formula

Consider two sets of situations, σ and σ', consisting of K situations each. If the variance of B_s is constant across all s in σ and σ', and if the correlation coefficients $r_{ss'}$ are the same for all pairs of situations within σ, within σ', and between σ and σ', then, according to the Spearman-Brown formula: $r_{\sigma\sigma'} = Kr_{ss'} / [1 + (K\text{-}1)r_{ss'}]$, where $r_{\sigma\sigma'}$ is the correlation coefficient between the averages Y_σ and $Y_{\sigma'}$ (e.g., Gulliksen 1950: 78). It follows that, no matter how small $r_{ss'}$ is, $r_{\sigma\sigma'}$ will approach 1 as K goes to infinity. Does it follow that we can get correlation coefficients as high as we want by aggregating behavior over sufficiently large numbers of situations?

Originally Epstein (1979a) did not appeal to the Spearman-Brown formula but gave (in addition to empirical evidence) only the intuitive argument I examined above. In fact, when Mischel and Peake (1982) disparaged Epstein's claim that aggregation increases correlation coefficients as nothing more than what would be expected from the Spearman-Brown formula, Epstein replied that the Spearman-Brown formula holds only "when standard assumptions are met that are rarely met in real-life situations";[121] "thus it is important to empirically investigate the effects of different forms of aggregation on the reliability and validity of real-life data" (1983: 180). Later on, however, Epstein himself used the Spearman-Brown formula (e.g., 1986: 1203), and claimed that, "given some degree of true relation to begin with, it is possible with sufficient aggregation to obtain a correlation of 1.00" (1986: 1204). In fact, Epstein has gone down in the literature as

---

[121] Epstein did not specify what these standard assumptions are, nor did he explain why they are rarely met in real-life situations.

"championing" the use of the Spearman-Brown formula (Ross & Nisbett 1991: 108); moreover, Ross and Nisbett claim that "in a sense, Epstein's argument was purely statistical and beyond dispute" (1991: 107).

While the truth of the Spearman-Brown formula is a mathematical fact and thus is indeed beyond dispute, I take the conclusion that with sufficient aggregation one can obtain a correlation coefficient of 1 to be a reductio of an objection (to my argument for the Independence Condition) that uses the Spearman-Brown formula. So I agree with Epstein's (1983a) original response to Mischel and Peake (1982): I take the Spearman-Brown formula to be inapplicable to many real-life situations because the assumption that all $r_{ss'}$ are equal is violated. To take a concrete example, suppose all situations in σ concern honesty in school situations, whereas all situations in σ' concern honesty in party situations. It is then reasonable to expect that $r_{ss'}$ for pairs of situations within σ or within σ' will be higher than $r_{ss'}$ for pairs of situations between σ and σ' (because the two situations in each pair of the former kind will be more similar to each other than the two situations in each pair of the latter kind will be similar to each other). Assuming, for example, that the former correlation coefficients are all equal to .3 and the latter are all equal to .1, one can compute $r_{\sigma\sigma'}$ by using the following formula, of which the Spearman-Brown formula is a special case: $r_{\sigma\sigma'} = Kr_{ss'\text{-between}} / [1 + (K\text{-}1)r_{ss'\text{-within}}]$ (e.g., Gulliksen 1950: 77). So as K goes to infinity $r_{\sigma\sigma'}$ goes to $r_{ss'\text{-between}} / r_{ss'\text{-within}} = .1/.3 = 1/3$ for my numerical example. One can thus go beyond Epstein's original vague appeal to the non-satisfaction of the assumptions behind the Spearman-Brown formula: the above considerations provide specific reasons for believing that the specific assumption of equality between correlation coefficients will fail and that this failure will invalidate the absurd conclusion that with sufficient aggregation one can obtain a correlation coefficient of 1.

It is finally worth noting that the above discussion was about prediction of *aggregate* behavior (over a number of situations) from a measure of *aggregate* behavior (over a number of other situations). When one tries to predict behavior in a *single* situa-

tion, using a measure of *aggregate* behavior would be of much less help even if the assumptions behind the Spearman-Brown formula were true. This is because, under these assumptions, the correlation coefficient $r_{s\sigma}$ between B_s and Y_σ (as opposed to the correlation coefficient $r_{\sigma\sigma'}$ between $Y_{\sigma'}$ and Y_σ) is given by the following formula: $r_{s\sigma} = K^{1/2}r_{ss'} / [1 + (K\text{-}1)r_{ss'}]^{1/2}$; thus $r_{s\sigma}$ goes to $r_{ss'}^{1/2}$, not to 1, as K goes to infinity. (E.g., for $r_{ss'} = .16$, $r_{s\sigma}$ goes to .4.)

4.4.3. Empirical evidence concerning aggregation

Although references to Hartshorne and May (1928, namely *Studies in the nature of deceit*, the first volume of *Studies in the nature of character*) are a staple of the debate about aggregation and cross-situational consistency, references to Hartshorne, May, and Shuttleworth (1930, namely *Studies in the organization of character*, the third volume of *Studies in the nature of character*) are much less frequent. But in the latter work the authors explicitly examine correlations between aggregate measures: they compute aggregate scores for (a) 23 deception (honesty) tests, (b) 5 cooperation tests, (c) 4 inhibition tests, and (d) 5 persistence tests, and then they compute correlation coefficients for pairs of these aggregate measures. These coefficients are again low, ranging from .049 to .361 (1930: 151). Epstein (like many others) neglects to mention these important results when he claims that "in almost all cases the correlations were based on single items of behavior" (1979a: 1102).[122]

---

[122] In their response to Epstein, Mischel and Peake (1982: 731) note: "far from overlooking reliability, virtually all of the classic, large-scale investigations of cross-situational consistency (e.g., Dudycha, 1936; Hartshorne & May, 1928; Newcomb, 1929) routinely employed behavioral measures aggregated over repeated occasions". Mischel and Peake (like many others), however, fail to note that (as I just said) Hartshorne, May, and Shuttleworth (1930) also employed behavioral measures aggregated over *situations* (not just over *occasions*).

<table>
<tr><td rowspan="6">(C1) Posterior probability of indeterminacy high</td><td rowspan="5">(L1) Prior probability of indeterminacy high</td><td rowspan="5">(L0) Most people are indeterminate</td><td rowspan="2">(Q1) Most people are fragmented</td><td>(Q3) In many situations most people (would) behave deplorably</td><td>- Milgram's obedience experiments
- Zimbardo's prison experiment
- Seizure experiments</td></tr>
<tr><td>(Q4) In many situations most people (would) behave admirably</td><td>- Theft experiments
- Electrocution experiments
- Rape experiments</td></tr>
<tr><td rowspan="3">(Q2) Fragmentation entails indeterminacy</td><td>(Q5) Deplorable behavior precludes goodness</td><td rowspan="3">(Q8) Every bad person is worse than every intermediate person
(Q9′) Every fragmented person is incommensurable with every paradigmatically intermediate person
(Q10) Every intermediate person is worse than every good person
(Q11) No fragmented person is worse than every good person</td></tr>
<tr><td>(Q6) Admirable behavior precludes badness</td></tr>
<tr><td>(Q7) Fragmentation precludes "intermediacy"</td></tr>
<tr><td colspan="2">(L2) Posterior probability of indeterminacy almost never much lower than prior</td><td colspan="2">(P1) If the Independence Condition holds, then posterior probability not much lower than prior</td><td>- Theorem 2.2</td></tr>
<tr><td colspan="2"></td><td colspan="2">(P2) The Independence Condition holds</td><td>- Low average correlations
- Personality variables don't predict help or obedience</td></tr>
</table>

Table 2.5. The *approximate* argument for the epistemic thesis

<table>
<tr><td rowspan="6">(C1) Posterior probability of indeterminacy high</td><td rowspan="5">(C1′) Posterior probability of fragmentation high</td><td rowspan="3">(L1′) Prior probability of fragmentation high</td><td rowspan="3">(Q1) Most people are fragmented</td><td>(Q3) In many situations most people (would) behave deplorably</td><td>- Milgram's obedience experiments
- Zimbardo's prison experiment
- Seizure experiments</td></tr>
<tr><td>(Q4) In many situations most people (would) behave admirably</td><td>- Theft experiments
- Electrocution experiments
- Rape experiments</td></tr>
<tr><td colspan="2">(P1′) If the Independence Condition holds, then posterior probability of fragmentation not much lower than prior</td><td>- Theorem 2.2</td></tr>
<tr><td>(L2′) Posterior probability of fragmentation almost never much lower than prior</td><td colspan="2">(P2) The Independence Condition holds</td><td>- Low average correlations
- Personality variables don't predict help or obedience</td></tr>
<tr><td colspan="2" rowspan="3">(Q2) Fragmentation entails indeterminacy</td><td colspan="2">(Q5) Deplorable behavior precludes goodness</td><td rowspan="3">(Q8) Every bad person is worse than every intermediate person
(Q9′) Every fragmented person is incommensurable with every paradigmatically intermediate person
(Q10) Every intermediate person is worse than every good person
(Q11) No fragmented person is worse than every good person</td></tr>
<tr><td colspan="2">(Q6) Admirable behavior precludes badness</td></tr>
<tr><td colspan="2">(Q7) Fragmentation precludes "intermediacy"</td></tr>
</table>

Table 2.6. The *real* argument for the epistemic thesis

This completes my defense of the Independence Condition and thus of the claim that (C1′) the posterior probability of fragmentation should almost always be high, and thus of the epistemic thesis itself. Tables 2.5 and 2.6 summarize the argument of §2-§4: Table 2.5 gives the argument as stated in §1 and Table 2.6 gives the *real* argument as stated in §4.1.

5. Objections to the epistemic thesis

In this section I examine three objections to the epistemic thesis: the objection that *comparative* character evaluations are warranted and thus non-comparative evaluations also are (§5.1), the objection that the thesis is too incredible to be true (§5.2), and finally the very different "objection" that the thesis is uninteresting because it is trivially true (§5.3).

5.1. The comparative evaluations objection

As we saw in §4.3, the values of cross-situational consistency correlation coefficients are typically so low that if we know that a given person behaved deplorably (or admirably) in a given situation we cannot reliably predict how this person will behave in any other situation. Ross and Nisbett (1991: 116-8), however, argue that the typical values of cross-situational consistency correlation coefficients, low as they are, warrant inferences of the following sort: if Harry behaved admirably and Tom behaved deplorably in a given situation, then we should be much *more* confident that Harry rather than Tom will behave admirably, and that Tom rather than Harry will behave deplorably, in any other situation. But then it seems that *comparative* character evaluations of the form "Harry is better than Tom" are often epistemically warranted, and thus so are *non*-comparative character evaluations: necessarily, Harry is good exactly if he is better than most other people.

The claim that comparative character evaluations are often epistemically warranted is ambiguous: it can be understood as *precluding* widespread incommensurability, or as *allowing* such incommensurability provided that we can often detect it. Consider the latter understanding first, and for the sake of argument let me grant that on that understanding the claim is true: for most pairs of people A and B, we know whether A is better than B, B is better than A, A and B are equally good (or bad), *or A and B are incommensurable*. For the sake of argument let me also grant that, necessarily, anyone who is better than most other people is good. It does not follow that anyone is good, because it does not follow that anyone is better than most other people: if incommensurability is widespread, then it is possible that everyone is incommensurable with most other people. The epistemic thesis emerges unscathed.

In response one might slightly modify the objection: one might claim that, necessarily, anyone who is better than most *of those with whom she is commensurable* is good, so that a person can be good despite being incommensurable with most other people. I reply that if everyone in the world who is not bad were to die today, it is not the case that those of the remaining people (bad) who are today better than most bad people would become good tomorrow; so it is false that, necessarily, anyone who is better than most of those with whom she is commensurable is good.

It seems then that the comparative evaluations objection can work only if the claim that comparative character evaluations are often epistemically warranted is understood as *precluding* widespread incommensurability (rather than as allowing such incommensurability provided that we can often detect it). But so understood how could this claim ever be supported by Ross and Nisbett's considerations? The claim that we should be much more confident that Harry rather than Tom will behave admirably in any given situation is compatible with the claim that Harry and Tom are incommensurable with each other because it is compatible with the possibility that Harry would behave admirably and Tom deplorably in an open list of situations but Tom would behave admi-

rably and Harry deplorably in another such open list. Of course one might deny that this possibility entails incommensurability, but then we would be back to the discussion of my general argument for incommensurability in §3.4 and §3.5: Ross and Nisbett's considerations provide no *new* objection to that argument. I conclude that the comparative evaluations objection fails.

A variant of the comparative evaluations objection concludes not that the epistemic thesis is false but rather that it is uninteresting (cf. §5.3) because comparative character evaluations are all we care about. I reply that if the epistemic thesis is true then comparative evaluations are also in a sense undermined because incommensurability is prevalent (regardless of whether it can be reliably detected):

Theorem 2.3. *Suppose that (1) anyone who is better than everyone who is bad is either intermediate or good, (2) anyone who is worse than everyone who is good is either intermediate or bad, (3) anyone who is intermediate is better than everyone who is bad and worse than everyone who is good, (4) anyone who is as good (or bad) as someone who is intermediate is also intermediate, and (5) the betterness relation is transitive. Then anyone who is indeterminate is incommensurable with (i.e., neither better than nor worse than nor as good or bad as) everyone who is intermediate.*[123]

From Theorem 2.3 it follows that if indeterminacy is prevalent then incommensurability also is.[124]

---

[123] *Proof.* Suppose, for reductio, that someone p_1 who is indeterminate is not incommensurable with someone p_2 who is intermediate. Then there are three possibilities. (a) p_1 is better than p_2. Then, from (3) and (5), p_1 is better than everyone who is bad, so from (1) p_1 is either intermediate or good. (b) p_1 is worse than p_2. Then, from (3) and (5), p_1 is worse than everyone who is good, so from (2) p_1 is either intermediate or bad. (c) p_1 is as good (or bad) as p_2. Then, from (4), p_1 is intermediate. In all three cases we reach a contradiction, given that p_1 is indeterminate (i.e., neither good nor bad nor intermediate). □

[124] If there are N people—and thus $N(N\text{-}1)/2$ pairs of people—and 70% of them are indeterminate (§2.4) whereas 15% of them are intermediate, then from Theorem 2.3 it follows that there are at least $(.7N)(.15N) \cong .10N^2$ incommensurable pairs, so that more than 20% of all pairs are incommensurable. If in addition many pairs of *indeterminate* people are incommensurable, then the percentage of incommensurable pairs can be much higher.

5.2. The incredibility objection

Experiments like Milgram's and Zimbardo's suggest that most people are *experimentally* fragmented: they (would) behave deplorably in many and admirably in many other experimental situations. Now even if *real-life* fragmentation suffices for indeterminacy it seems incredible to suggest that experimental fragmentation does. But then to conclude that most people are indeterminate is to commit the fallacy of equivocation, to conflate two senses of 'fragmentation'—or so this second objection to the epistemic thesis goes. In reply I will argue that we do have evidence for real-life fragmentation, and that in any case even experimental fragmentation suffices for indeterminacy.

Why can't we infer from how people behave in experiments how they (would) behave in real-life situations? Because, one might argue, in experiments people (a) *volunteer* or at least *consent* to place themselves in (b) *artificial* situations in which (c) they are *aware* of being observed. Now these considerations are clearly inapplicable to my evidence for the prevalence of *admirable* behavior (§2.3): the theft, electrocution, and rape experiments simulated real-life situations, and the participants did not know that an experiment was taking place. So the objection applies only to my evidence for the prevalence of *deplorable* behavior (§2.2), namely the obedience, prison, and seizure experiments. I reply that there are also more naturalistic studies in which people behave deplorably. In an experiment by Hofling, Brotzman, Darlymple, Graves, and Pierce (1966), nurses at two hospitals received a telephone call from an experimenter who identified himself as a physician and asked each nurse to administer an obviously excessive dose of an unfamiliar medicine to a patient; 21 of 22 nurses complied (they were stopped by another experimenter), in violation of hospital policy against telephone medi-

cation orders.[125] In another experiment (West, Gunn, & Chernicky 1975), inspired by Watergate, undergraduate criminology majors were approached by an experimenter who identified himself as a government agent and presented them with elaborate plans for burglarizing a local advertising firm in order to microfilm an allegedly illegal set of accounting records maintained by the firm to defraud the U. S. government out of 6.8 million tax dollars per year; 9 of 20 participants who were guaranteed immunity from prosecution if apprehended agreed to commit the burglary (whereas one of 20 participants who were warned that there would be no immunity agreed). More generally, there is informal evidence that the phenomenon of excessive obedience to authority is not restricted to Milgram's laboratories. Tarnow (2000: 120) estimates that an important factor in as many as 25% of all airplane accidents is excessive obedience by first officers to captains' erroneous orders. Browning (1992) describes how middle-aged reserve German policemen ("ordinary men") shot some 1,500 Jews in a Polish village in the summer of 1942. Recent events in Rwanda (Gourevitch 1998) and other countries suggest that under certain circumstances most ordinary people will inexcusably commit multiple murders. Some people may find such evidence less convincing than controlled experiments (while others may find it more convincing), but in any case the evidence does pertain to real-life situations.

In response one might argue that not *every* real-life situation is relevant to character evaluations: only *everyday*-life situations are. The idea is that we evaluate (e.g.) our friends as "good people" on the basis of how they behave in everyday life, not how they would behave in extraordinary situations like those in Rwanda or World War II. In reply I ask: why is deplorable (or admirable) behavior in "extraordinary" situations not sup-

---

[125] Rank and Jacobson (1977) report a "failure to replicate", but their study differed from Hofling et al.'s in several respects; e.g., the nurses were familiar with the drug, had the opportunity to interact with other nurses, and had volunteered a few days in advance to participate in an experiment (whose nature and time had not been disclosed).

posed to count? One might argue that wrong behavior in such situations is adequately excused. But this is not *always* so (one can behave inexcusably even in wars), and when it is *not* so why shouldn't the behavior count? (When the behavior *is* adequately excused then by definition it is not deplorable: see Objection 3 in §3.3.2.) One might also argue that in extraordinary situations people may behave "out of character" (cf. Hampshire 1953: 7-8). But even out-of-character behavior counts if it is *seriously* blameworthy (as deplorable behavior by definition is): your vicious murder may be mitigated but is not adequately excused by your being ordinarily a model citizen. (Moreover, can one behave "out of character" in an *open list* of situations?) More generally, the idea that only every-day-life behavior is relevant to character evaluations seems misguided: the relevance of behavior in a given situation to character evaluations depends not on how ordinary or extraordinary the situation is but on how trivial or significant the behavior is. Wars and plagues may be extraordinary, but behavior in them can be revealing in ways in which habitual behavior in everyday life is not (cf. Kupperman 1991: 160). Deplorable behavior is never trivial (because by definition it is *seriously* blameworthy), so it is nonmarginally relevant to character evaluations regardless of whether it occurs in a pedestrian or an outlandish—even *experimental*—situation. These general considerations suggest that experimental fragmentation suffices for indeterminacy; real-life fragmentation is not needed. So I am *not* using the claim that a person would behave deplorably in many experimental situations to infer that the person would also behave deplorably in many real-life situations and thus is not good; I am rather arguing that, since a good person would *not* behave deplorably in many situations (experimental or not), the fact that a person *would* behave deplorably in many experimental situations *disconfirms* the hypothesis that the person is good (cf. Mook 1983: 383).

But again, isn't the claim that experimental fragmentation suffices for indeterminacy simply incredible? Take your favorite case of a good person; for example, your mother. Suppose you have observed your mother's behavior over many years in widely

varied situations and she has never behaved deplorably. Isn't it incredible to deny that she is good because she *would* behave sadistically in (e.g.) Zimbardo's experiment? In reply consider the precursor of Zimbardo's experiment which was carried out by a group of undergraduates (§2.2.2):

> They divided themselves into prisoners and guards ... [T]heir experience was profound... By the end of the weekend some long-term friendships were broken because those young men and women who were prisoners believed that in their roles as "mock" guards the "true" self of their former friends was revealed, and they could no longer befriend such sadistic authoritarian people (Zimbardo 1975: 37).

These people *knew* it was an experiment; did this make it irrelevant for them? One might object that in this case the deplorable behavior *actually* occurred. In reply consider a questionnaire item from my study:

> Suppose you were to learn (never mind *how*) for *certain* that in a situation which is *very unlikely* to arise (e.g., a flood, plague, war, or a strange psychological experiment) your best friend would behave *very badly* towards you; e.g., (s)he would refuse to help you although (s)he could help you with little effort. (Assume that the severity of the situation *would not sufficiently excuse* your friend's behavior.) Does the fact that your friend *would* behave like this (although (s)he very probably *won't*, since the situation is very unlikely to arise) count as relevant to your assessment of your friend's moral character?

About 81% (i.e., 26) of 32 students answered "yes", and about half of those who did so said that the extraordinary, counterfactual behavior counts as *more* relevant "than the fact that in everyday life your friend always behaves admirably (e.g., is always nice and goes out of her/his way to help you and other people, never breaks promises, does volunteer work for charities, etc.)". So my claim that even experimental fragmentation suffices for indeterminacy is not so incredible after all.

5.3. The triviality objection

The triviality objection claims that the epistemic thesis is true but trivial: we already *know* that character evaluations are epistemically unwarranted, so we seldom make such evaluations. Talk of character occurs mostly in special contexts like recommendation letters; in everyday discourse we only rarely use expressions like "he's a man of good character"—or so the objection goes. I have three replies.

First, even if we seldom *express* character evaluations, we may still often *make* such evaluations and refrain from expressing them because expressing them is socially undesirable: telling you to your face that you are bad or good is insulting or ingratiating respectively, and telling you that a third person is good or bad is often in bad taste.

Second, there is reason to believe that we do often make character evaluations. The concept of character is a commonplace, not an arcane or an obsolete one, and we often have attitudes which *presuppose* character evaluations: I esteem you when I evaluate you as good, and I despise you when I evaluate you as bad. It's true that, as we saw in Chapter I, esteem (or appraisal respect) is sometimes partial or even non-moral. But the fact that esteem is sometimes (or even often) partial or non-moral is compatible with my claim that esteem is often global and moral—i.e., based on an evaluation of a person's character as good.

Third, the question of how often we make character evaluations is an empirical one, so I collected some relevant data. I asked 33 introductory psychology students questions 1 and 2, and 32 other introductory psychology students questions 3 and 4:

> 1. How often does it happen that you evaluate someone *in conversation* in terms of their moral character (in other words, you *say to someone* something *like* "she is a good [or bad] person")?
>
> 2. How often does it happen that you evaluate someone *in your mind* in terms of their moral character (in other words, you *say to yourself* something *like* "she is a good [or bad] person", but you *don't* go on to voice this thought)?
>
> 3. How often does it happen that people evaluate someone *in conversation* in terms of moral character? In other words, how often do you *hear people saying* something *like* "she is a good [or bad] person"?

4. How often does it happen that people evaluate someone *in their mind* in terms of moral character? In other words, how often do people *say to themselves* something *like* "she is a good [or bad] person", but they *don't* go on to voice this thought?

The results, reported in Table 2.7, suggest that character evaluations both are and are considered to be common, especially when they are not expressed. For example, the modal response to question 2 was "somewhat frequently", and the mean response was between "somewhat frequently" and "frequently". I conclude that the triviality objection is implausible.

Question #	1 Never	2 Almost never	3 Rarely	4 Somewhat rarely	5 Somewhat frequently	6 Frequently	7 Very frequently	8 Almost always	Mean response
1	0	8	4	**10**	6	3	1	1	4.45
2	0	0	4	3	**14**	6	2	4	5.33
3	0	1	3	3	**11**	7	6	1	5.31
4	0	0	0	1	7	8	**10**	6	6.40

Table 2.7. Data pertaining to the triviality objection

6. Related work by John Doris and Gilbert Harman.

The epistemic thesis represents a reaction to a set of surprising psychological results. These results exemplify a long-standing *situationist* research tradition in social and personality psychology, a tradition whose central tenet I take to be that the behavior of a given person in a given situation depends more on characteristics of the situation and less on characteristics of the person than people typically assume. In recent years a small but growing body of philosophical literature (e.g., Athanassoulis 2000; Bok 1996; Campbell 1999; Cullity 1995; DePaul 2000; Doris 1996, 1998, in press; Flanagan 1991; Harman 1999, 2000; Kupperman 2001; Merritt 1999, 2000; Pigden & Gillet 1996; Railton 1995; Sreenivasan 2000) has emerged as a reaction to situationist results, especially

after the publication of Ross and Nisbett's summary of such results in *The person and the situation* (1991). Doris and Harman, in particular, have proposed arguments apparently similar to my argument for the epistemic thesis, so I would like to compare their arguments with mine.

Recall that, according to my argument for the prevalence of indeterminacy, most people cannot be evaluated in terms of moral character because character evaluations presuppose behavioral consistency (i.e., Q2: fragmentation entails indeterminacy) but most people lack consistency (i.e., Q1: most people are fragmented). Doris puts forward an apparently similar line of reasoning: "trait attribution requires substantial *cross-situational consistency* in behavior … [but] systematic observation typically reveals failures of cross-situational consistency" (1998: 507). Although Doris is talking here about *trait attributions*, apparently he endorses an analogous line of reasoning about *character evaluations*, "global personality judgments like 'good person' or 'bad person' " (1998: 514). Doris concludes that "people typically lack character" (1998: 506), and Harman similarly concludes that "there is no evidence that people have character traits" (1999: 315). Harman (2000: 225) explicitly contrasts his conclusion, which is about "people" in general, with a conclusion—like the one of my argument—which is about *most* people. Harman, however, would presumably accept—and Doris (1998: 524 n. 33) explicitly accepts—that situationist results are consistent with the *possibility* that a minority of people do have character traits (or are not indeterminate); so on a charitable reading Harman's point is that there is *no evidence* that this possibility is actual.[126] Doris (1998: 511) suggests that this possible minority is small, but in §2.4 I derived an approximate lower bound on the percentage of fragmented people: 69%, as it turned out. But then the psy-

---

[126] Harman (2000: 225) notes that "in Milgram (1963) *every* subject was willing to apply shocks of up to 300 volts". But it does not follow that *everyone* is fragmented: see footnote 18 and the beginning of §2.4. (Contrary to the obedience experiment I described in §2.2.1, in which the learner explicitly withdrew his consent at 150 volts, in the obedience experiment to whom Harman refers "no vocal response or other sign of protest is heard from the learner until Shock Level 300 is reached" (Milgram 1963: 374).)

chological results are consistent with the possibility that a *sizeable* minority (up to about 31%) of people are *not* fragmented, and it seems prudent to keep this possibility in mind rather than dismissing it on the ground that there is no evidence for its actuality.

Although Doris's argument may look similar to mine, there are two reasons why it is in fact significantly different. (1) Doris apparently understands cross-situational inconsistency not as what I call 'fragmentation', but rather as *actual* (not merely counterfactual) *high variability* in a person's behavior across a range of relevant situations. This kind of inconsistency, however, does not entail (fragmentation or) indeterminacy: a person who never behaves deplorably (and thus is not fragmented) but often behaves neutrally and often *extremely* admirably can be good (rather than indeterminate) but also inconsistent in Doris's sense. But then Doris's argument is unsound (similarly for Harman's). (2) To support his claim that people typically lack behavioral consistency, Doris (1998) appeals not to experiments (like Milgram's and Zimbardo's) in which most people behave deplorably or admirably, but rather to studies and reviews (like those I used in §4.3) which suggest that the average correlation coefficients between people's behaviors in various situations are typically small.[127] I will argue now, however, that the evidence for low average correlation coefficients fails to show that most people are fragmented or even inconsistent in Doris's *own* sense; at most it shows that most people lack a kind of consistency which is *not* presupposed by character evaluations.

Consider a numerical measure X_{ps} of how person p (p = 1, …, P) behaves in situation s (s = 1,…, S); e.g., X_{ps} can be -1, 0, or 1 if p behaves in s deplorably, neutrally, or admirably respectively. The correlation coefficient $r_{ss'}$ is defined as $\Sigma_p Z_{ps} Z_{ps'}/P$, where Z_{ps} is the *standardized* measure of p's behavior in s, *relative* to how the other P-1 people

---

[127] Doris (1996, in press) does also appeal to Milgram's and Zimbardo's experiments, but to support the *different* claim that "[b]ehavioral variation across a population owes more to situational differences than dispositional differences among persons" (1998: 507). Harman also appeals to Milgram's experiment, but to illustrate the "fundamental attribution error" (2000: 223; cf. footnote 51).

behave in s (think of Z_{ps} as a "percentile score").[128] Given that $r_{ss'}$ is defined in terms of *everyone's* behavior in *two* situations (s and s'), how can $r_{ss'}$ be linked to a measure of cross-situational (in)consistency like $V(Z_p)$—the variance of p's standardized behavior across the S situations—which is defined in terms of a *specific* person's behavior in *all* situations? An answer is provided neither by Doris nor by Harman nor by the psychological literature on cross-situational consistency, but rather by the following theorem:

Theorem 2.4. $\frac{1}{S(S-1)}\sum_{s}\sum_{s'\neq s} r_{ss'} = 1-\frac{S}{S-1}\frac{1}{P}\sum_{p}V(Z_p)$.[129]

In words: the average correlation coefficient (across all $S(S\text{-}1)$ ordered pairs of situations) is inversely related to the average variance of standardized behavior (across all P persons). So the lower the average correlation coefficient, the higher the average variance of standardized behavior, and thus the lower the average cross-situational consistency.

Given Theorem 2.4 and the evidence for low average correlation coefficients, let me grant that most people are behaviorally inconsistent in the sense of being *Z-fragmented*: having high $V(Z_p)$. Being Z-fragmented, however, entails neither being *X-fragmented* (having high $V(X_p)$, which is apparently how Doris and Harman understand behavioral inconsistency) nor being fragmented (in *my* sense) nor being indeterminate. To see this consider a person who always behaves in the same way (and thus is not X-fragmented: $V(X_p) = 0$), namely neutrally (and thus is not fragmented: she never behaves

---

[128] But note that Z_{ps} is defined as $(X_{ps}\text{-}E(X_s))/\sigma(X_s)$ and thus can be negative. X_s is the "column" variable $(X_{1s} \ldots X_{Ps})^T$ (and X_p is the "row" variable $(X_{p1} \ldots X_{pS})$). $E(X_s) = \Sigma_p X_{ps}/P$ and $\sigma^2(X_s) = \Sigma_p(X_{ps} - E(X_s))^2/P = V(X_s)$.

[129] *Proof.* By definition, $V(Z_p) = \Sigma_s Z_{ps}^2/S - (\Sigma_s Z_{ps}/S)^2$, so $\Sigma_p V(Z_p)/P = \Sigma_p\Sigma_s Z_{ps}^2/SP - \Sigma_p(\Sigma_s\Sigma_{s'} Z_{ps}Z_{ps'})/S^2P = \Sigma_s(\Sigma_p Z_{ps}^2/P)/S - \Sigma_s\Sigma_{s'}(\Sigma_p Z_{ps}Z_{ps'}/P)/S^2$. Given that $\Sigma_p Z_{ps}^2/P = V(Z_s) = 1$ and $\Sigma_p Z_{ps}Z_{ps'}/P = r_{ss'}$, we get: $\Sigma_p V(Z_p)/P = 1 - \Sigma_s\Sigma_{s'} r_{ss'}/S^2 = 1 - (S + \Sigma_s\Sigma_{s'\neq s} r_{ss'})/S^2$, from which Theorem 2.4 easily follows. □

deplorably or admirably). Such a person can be intermediate (between good and bad) rather than indeterminate but can still be *Z*-fragmented; for example, she can disobey relatively early in Milgram's experiment (and thus behave better than average: high Z_{ps}), alert the security officer in Harari et al.'s rape experiment (and thus behave worse than average: low Z_{ps}), and so on. Moreover, even if Doris or Harman were somehow to link the evidence for low average correlation coefficients with *X*-fragmentation,[130] the problem would remain that, as I argued above, even *X*-fragmentation does not entail indeterminacy or fragmentation.

I conclude that (Doris's and Harman's appeal to) the evidence for low average correlation coefficients fails to establish that most people are fragmented or that they lack a kind of behavioral consistency which is presupposed by character (or even trait) evaluations. More work is required than Doris and Harman undertake if we are to accept that most people are indeterminate or lack character traits. In this chapter I carried out the extra work on indeterminacy; I did not take a stand on character traits.

7. The pragmatic thesis

All right. Suppose I have convinced you that you and your loved ones are probably indeterminate. You may still try to hold on to your cherished character evaluations on the ground that doing so would be overall most beneficial: how could you keep loving your mother if you regarded her as "fragmented", how could you maintain your self-esteem if you viewed yourself as "indeterminate"? This move is in a way analogous to Pascal's wager: we have good *pragmatic* reason to act so as to become (or keep being)

---

[130] I don't see how this could be done. A reasoning similar to that in footnote 129 for X_p in the place of Z_p gives: $\Sigma_p V(X_p)/P = \Sigma_s(\Sigma_p X_{ps}^2/P)/S - \Sigma_s\Sigma_{s'}(\Sigma_p X_{ps}X_{ps'}/P)/S^2 = \Sigma_s(V(X_s)+(E(X_s))^2)/S - \Sigma_s\Sigma_{s'}(E(X_s)E(X_{s'})+ r_{ss'}\sigma(X_s)\sigma(X_{s'}))/S^2$. So we need further assumptions if we are to generalize Theorem 2.4 so as to relate $\Sigma_s\Sigma_{s'\neq s} r_{ss'}/S(S\text{-}1)$ with $\Sigma_p V(X_p)/P$ (as opposed to $\Sigma_p V(Z_p)/P$).

theists even if theism is epistemically unwarranted.[131] In response I introduce my preferred alternative to character evaluations: *local* evaluations of people in light of their behavior in relatively restricted ranges of actual and counterfactual situations. You are probably indeterminate; still, I may be to some extent epistemically justified in evaluating you as good *insofar* as you are regularly nice to your colleagues. In evaluating you thus I keep in mind that I cannot confidently predict how you would behave in situations other than those routine interactions with your colleagues in which I have already observed your behavior.[132] Of course to the extent that your behavior even in relatively specific situations can still vary widely according to, e.g., your mood, even local evaluations may not be epistemically justified; but they will normally be *less unjustified* than character evaluations. My *pragmatic* thesis is that we have good pragmatic reason to *prefer* local to character evaluations (so that character evaluations are pragmatically unwarranted: they are not the overall most beneficial alternative). I wish to conclude this chapter by very briefly sketching an argument for the pragmatic thesis.

My argument for the pragmatic thesis relies on two considerations. (1) By keeping people's fragmentation salient in our minds, local—in contrast to character—evaluations help us avoid creating situations in which people (ourselves included) will show their dark sides, and help us create situations in which they will show their bright ones. For example, if I realize that I cannot confidently predict my behavior in situations I have never encountered, I may be inclined to avoid morally dangerous situations rather

---

[131] One important disanalogy is that Pascal's wager is concerned with costs and benefits *only* to the—potential—believer.

[132] Note that evaluations like "good colleague" or "good spouse" don't correspond to what I call *local* evaluations because they presuppose significant counterfactual behavioral stability: a good spouse, for example, is expected to behave in a certain way even if, e.g., the other spouse becomes disabled. It is an open question whether we should reinterpret expressions like "good spouse" so as to make them refer to local evaluations or whether we should introduce new terms for local evaluations.

than facing them with the misplaced confidence that I will overcome temptation.[133] As another example, if I realize that my spouse may behave admirably towards me only as long as the circumstances are propitious, I may be inclined to keep the circumstances propitious rather than subjecting her love to "tests" which may result in friction and disappointment. (2) Character evaluations have useful functions: they help us regulate our emotions towards people and decide whom to avoid and whom to associate with. But these benefits can be reaped by local evaluations almost equally well: you can keep loving your mother if you evaluate her as good *in ways that matter* (e.g., for your interaction with her).

---

[133] It might be objected that if I never take risks, if I never face temptations, then I will be unprepared when temptations come up (as they unavoidably will). I reply that I can take risks provided that I do so progressively and with caution.

CHAPTER III
A PARTIAL PRAGMATIC DEFENSE OF RECOGNITION RESPECT

In Chapter II I presented an epistemic *critique* of *one* of the two basic kinds of respect distinguished in Chapter I: appraisal respect. In this chapter I present a partial pragmatic *defense* of the *other* basic kind: recognition respect. I argue that we have strong prudential reason to avoid a central kind of disrespectful behavior: insulting, belittling, humiliating, ridiculing, and so on. Three insights underlie my argument. (i) Disrespect often leads to retaliation spirals. (ii) Disrespect is "addictive": strategically insulting when it is advantageous makes one more likely to insult even when it is not. (iii) Respect is a skill: practice makes one increasingly efficient at finding and implementing nondisrespectful ways of achieving one's goals.

In §1 I introduce my thesis and my argument. In §2 I formulate my thesis in detail. In §3 I argue that consistently behaving nondisrespectfully is better than consistently behaving disrespectfully. In §4 I conclude by arguing very briefly that behaving nondisrespectfully is also better than three other competing strategies.[1]

---

[1] Material from this chapter was presented at the seventy-fifth annual meeting of the Pacific Division of the American Philosophical Association (San Francisco, March 2001) and at the University of Michigan (Decision Consortium seminar, April 1999, April 2000, and March 2001). In addition to the debts I mentioned in the Acknowledgments, I am indebted to the following people for interesting questions: Michael Bratman, Lee Green, Eric Gampel, Christopher Hitchcock, Aviv Hoffmann, Frances Kamm, Markus Kemmelmeier, Don Loeb, Dominic Murphy, Ruth Sample, Tony Smith, and Frank Yates.

1. Introduction

This is a tough world. People who are usually all sweetness and light can turn very nasty when conflicts emerge. They can deny us what we deserve; they can trample on our rights; they can slander, threaten, or insult us. Such incidents are more frequent in some environments than in others: in academia they may be rare, among mafiosi they may be common. But none of us is immune from them, and when they occur it seems that we can avoid being exploited only by turning nasty ourselves. In some settings we may need to do so only *in reciprocation*; in others we may also need to do so *preventively*, to establish or maintain a reputation as a person not to be messed with. Exactly when to turn nasty is a delicate matter, but this much seems clear: we cannot afford to be invariably respectful. And yet I will argue that this commonsensical observation is mistaken: avoiding disrespect is prudentially rational.

The word 'disrespect' can be understood *broadly*, as including almost every transgression of the moral law: killing evinces disrespect towards human life, stealing evinces disrespect towards property rights, and so on. I will understand 'disrespect' much more narrowly, as limited to behavior like insulting, belittling, humiliating, ridiculing, and so on. It follows that the prescription to *avoid* disrespect leaves plenty of leeway about what to *do*; more technically, the strategy of avoiding disrespect is *coarse-grained*. This observation raises two worries. First, avoiding disrespect is consistent with acting immorally: just as it is possible to kill or steal disrespectfully, it is possible to kill or steal nondisrespectfully. So am I saying that killing and stealing are fine as long as you don't insult your victims? No. I am rather saying that justifying the prescription to avoid disrespect in this narrow sense does not amount to justifying the whole of morality. In particular, it does not amount to justifying the prescriptions to avoid killing and stealing. It amounts only to adding a pebble to what Kavka (1984) called the "Reconciliation Project": the project of reconciling the apparently conflicting demands of morality and rationality, of showing that it is rational to behave morally. Now a sec-

ond worry is that avoiding disrespect is consistent not just with *immoral* behavior like killing and stealing, but also with *imprudent* or *ineffective* behavior like turning the other cheek or meekly swallowing insults; so how can avoiding disrespect be prudentially rational? To see how it can, take an analogy. One can avoid smoking consistently with engaging in unhealthy behavior like eating fatty foods or even taking drugs. Still, the prescription to avoid smoking is certainly justified from the point of view of health promotion. The point is that the *healthiest* fully specified ways of living will not include smoking; similarly, my thesis is that the *most effective* fine-grained strategies, those which *best* promote our welfare, will not include disrespect. This is the sense in which I claim that we have strong prudential reason to follow the coarse-grained strategy of avoiding disrespect. Call this strategy for short *habitual effective nondisrespect* (HEN). The word 'we' in the formulation of my thesis is intentionally imprecise: it refers to people like us but allows for exceptions. Just as there are possible or maybe even actual people whose health is not harmed by smoking, there are possible or maybe even actual people whose self-interest is not harmed by behaving disrespectfully; in neither case does the existence of exceptions make the corresponding thesis uninteresting.

I will not argue that HEN is better than *every* other possible (long-term) strategy; I will argue rather that it is better (i.e., has higher long-term expected utility) than the following four initially appealing (kinds of) strategies. (1) *Habitual effective disrespect* (HED), which prescribes consistent disrespectful behavior. (2) *Case-by-case decisions* (CBC), which prescribes deciding on a case-by-case basis whether to behave disrespectfully or nondisrespectfully. (3) *Indicator-differentiated habits* (IDH), which prescribes behaving disrespectfully or nondisrespectfully depending on whether one detects the presence or the absence of a *reliable indicator* of the fact that disrespect is in one's short-term interest. (4) *Tit for tat* (TFT), which prescribes (a) behaving nondisrespectfully at the first move of an interaction (if the first move is yours) and (b) behaving disrespectfully at each of your other moves exactly if your interaction partner has behaved disre-

spectfully at the previous move. The bulk of this chapter is devoted to defending the superiority of HEN to HED; in §4 I explain how my argument can be extended to defend the superiority of HEN to CBC, IDH, and TFT. But before I defend HEN (§3) I want to explain in more detail what exactly HEN prescribes (§2.1) and in what sense I claim that HEN is rationally warranted (§2.2).

2. The thesis in more detail

2.1. What exactly does HEN prescribe?

2.1.1. Insulting behavior

In §1 I said that HEN prescribes avoiding only a specific kind of disrespectful behavior: insulting, belittling, humiliating, ridiculing, and so on. This was very rough, however; more precisely, HEN prescribes avoiding behavior *likely to be interpreted as insulting* (belittling, and so on). Now what exactly is insulting behavior? Let me first make clear that asking whether a behavior *is insulting* differs from asking whether anyone *feels insulted* by the behavior and from asking whether the behavior *is an insult*. (Saying that a behavior *insulted* a person may mean that the behavior was insulting, that the person felt insulted by the behavior, or both.) Whether a behavior is insulting is determined (I will argue) by how the behavior is intended; whether anyone feels insulted by a behavior is determined by how the behavior is interpreted; and whether a behavior is an insult is determined by the conventions that apply in the context of the behavior. (For example, "in certain tribal cultures in Nigeria, to raise one's hand toward another, much as Americans would wave 'Hi!' in greeting, is considered a very serious insult" (Flynn 1977: 13).) Here is why it is neither necessary nor sufficient for a behavior's being insulting that (1) anyone feel insulted by the behavior or that (2) the behavior be an insult. (1a) A behavior may be insulting even if nobody feels insulted by it: you may fail to notice my rudeness because you are distracted, or you may fail to understand my obscenity because you are a foreigner. (1b) One may feel insulted by a behavior which is not in-

sulting: you may misinterpret my innocuous remark because you are in a bad mood. (2a) A behavior may be insulting without being an insult: directing to you a badly mispronounced obscenity (e.g., because I don't have the guts to pronounce the obscenity correctly, or because I am a foreigner with a limited grasp of English) is insulting and may even be recognized by you as such.[2] (A badly mispronounced obscenity need not be an insult: there is no convention according to which /mezefu:ke/—as opposed to /mʌðərfʌkər/—is offensive.)[3] (2b) A behavior may be an insult without being insulting: consider teasing (in the sense of "bantering"). "If people are close enough to one another that they are able to kid each other in an insulting way, an exchange of insults may serve to both define and sustain their primary bonds. The use of insult as an indirect mode of expressing friendship and affection is particularly common among males, for whom the direct expression of friendship is considered awkward" (Flynn 1977: 82). One may object that, even though in friendly contexts teasing is not insulting, in such contexts teasing is not an insult either, because the general convention that makes teasing an insult is overriden in such contexts by a special convention that makes teasing an expression of friendship.[4] I reply that no such special convention exists the *first* time one feels suffi-

---

[2] It seems necessary that my remark be at least barely recognizable as an obscenity: directing to you a perfectly polite remark which I mistakenly take to be an obscenity may not be insulting. It seems thus that, in order for a behavior to be insulting, the intention that the behavior be interpreted in a certain way should not be based on totally unreasonable beliefs; I am assuming that this condition is satisfied.

[3] Compare my claim that a behavior may be insulting without being an insult to the claim that communication is possible in the absence of a code. "For example, Peter asks Mary, ['How are you feeling today?'] Mary responds by pulling a bottle of aspirin from her bag and showing it to him. Her behavior is not coded: there is no rule or convention which says that displaying a bottle of aspirin means that one is not feeling well. Similarly, her behavior affords only the weakest kind of direct evidence about her feelings: maybe she always carries a bottle of aspirin in her bag. On the other hand, it is strong direct evidence of her intention to inform Peter that she does not feel well. Because her behavior enables Peter to recognize her intention, Mary successfully communicates with him, and does so without the use of any code" (Sperber & Wilson 1995: 25-6).

[4] I agree that such a special convention exists in the case of actors uttering on the stage what are normally considered to be insults; therefore, actors on the stage normally provide no example of noninsulting behavior which is an insult.

ciently comfortable with someone to try out teasing.[5]

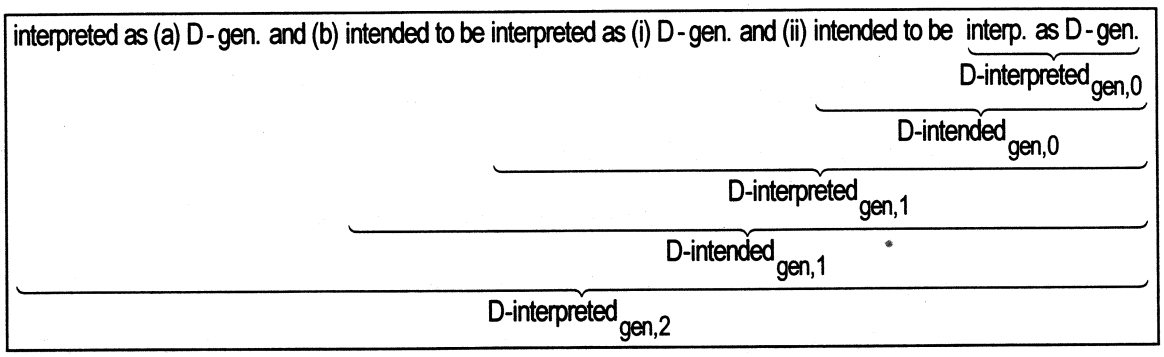
I conclude that whether a behavior is insulting is determined neither by how the behavior is interpreted nor by what conventions are in place. This conclusion supports my claim that whether a behavior is insulting is determined by how the behavior is intended. But what kind of intention makes a behavior insulting (i.e., what is an "insulting intention")? In order to even formulate my answer, I need to introduce some terminology.

2.1.2. A vocabulary for (dis)respect-related behaviors

The terminology that I will introduce is of general interest: it enables one to describe precisely almost any kind of (dis)respect-related behavior. My vocabulary consists of compound adjectives (e.g., 'D-generated'), the first part of each adjective being a prefix (e.g., 'D' for 'disrespect'; more generally, the prefix stands for one of the attitudes of respect and disrespect that I identified in Chapter I). There are four main classes of adjectives in my vocabulary: I will define 'D-generated', 'D-appearing', and various kinds of 'D-interpreted' and of 'D-intended' behavior. Here are the definitions. (1) A behavior is *D-generated* exactly if the source's (attitude of) D (disrespect) is causally implicated in the generation of the behavior. (The *source* is the agent who performs the behavior.) (2) A behavior is *D-appearing* exactly if there is a convention according to which the behavior expresses D. (3_0) A behavior is *D-interpreted*$_{gen,0}$ exactly if it is interpreted as being D-generated. ($4_{\mathbf{N}}$) For any $i \in \mathbf{N} = \{0, 1, 2, \ldots\}$, a behavior is *D-intended*$_{gen,i}$ exactly if it is intended to be D-interpreted$_{gen,i}$. ($3_{\mathbf{N}^*}$) For any $i \in \mathbf{N}^* = \{1, 2, \ldots\}$, a behavior is *D-interpreted*$_{gen,i}$ exactly if it is interpreted as being both D-generated and D-

[5] Of course "the object of an intended joke [may feel] that his incipient friendship with the teaser was not strong enough at that point to justify defining the action as just kidding. Thus, any kind of kidding or non-maliciously intended insult behavior, even among close friends, may well backfire" (Flynn 1977: 84). Nevertheless, as I argued in 1b, the fact that someone feels insulted by a behavior does not make the behavior insulting.

intended$_{gen,i-1}$. (Similarly, for any $i \in \mathbf{N}$, *D-interpreted*$_{app,i}$ and *D-intended*$_{app,i}$ behavior are defined with respect to D-appearing rather than D-generated behavior.[6]) Note the recursive structure of definitions $4_{\mathbf{N}}$ and $3_{\mathbf{N}^*}$: D-intended$_{gen,0}$ behavior is defined in terms of D-interpreted$_{gen,0}$ behavior, D-interpreted$_{gen,1}$ in terms of D-intended$_{gen,0}$, D-intended$_{gen,1}$ in terms of D-interpreted$_{gen,1}$, D-interpreted$_{gen,2}$ in terms of D-intended$_{gen,1}$, D-intended$_{gen,2}$ in terms of D-interpreted$_{gen,2}$, and so on. Here is a fully spelled out example. A behavior is D-interpreted$_{gen,2}$ exactly if it is:



Equivalently,[7] a behavior is D-interpreted$_{gen,2}$ exactly if it is interpreted as being: (0) D-generated; (1) intended to be interpreted as being D-generated; and (2) intended to be interpreted as being intended to be interpreted as being D-generated.[8] Note that, for any $i \in \mathbf{N}^*$, if a behavior is D-interpreted$_{gen,i}$ then it is D-interpreted$_{gen,i-1}$, and if a behavior is D-intended$_{gen,i}$ then it is D-intended$_{gen,i-1}$. I will say that a behavior is *D-*

---

[6] One can also define *D-intended*$_{app}$ behavior as behavior intended to be D-appearing. (Behavior intended to be D-generated is of less interest.)

[7] I assume that "interpreted as being both x and y" is equivalent to "interpreted as being x and interpreted as being y", and that "intended to be interpreted as being both x and y" is equivalent to "intended to be interpreted as being x and intended to be interpreted as being y".

[8] Expressions like "intended to be interpreted as being intended to be interpreted as being D-generated" are elliptical for "intended by the source to be interpreted by t as being intended by the source to be interpreted by t' as being D-generated". I assume that $t = t'$; this assumption could be relaxed.

interpreted$_{gen,\infty}$ exactly if it is D-interpreted$_{gen,i}$ for every $i \in \mathbf{N}$; similarly for *D-intended*$_{gen,\infty}$ behavior.[9]

Some clarifications and comments are in order. (a) Different conventions apply to different (groups of) people; thus a behavior can be D-appearing *to* some but not all members of an audience. Similarly, a behavior can be interpreted in different ways by different people and can even be intended to be interpreted in different ways by different people;[10] thus a behavior can be, e.g., D-interpreted$_{gen,1}$ *by* (and D-intended$_{gen,1}$ *for*) some people but not others. (b) Interpretations are not always accurate; thus, e.g., D-interpreted$_{gen,0}$ behavior need not be D-generated. Similarly, appearances can mislead (D-appearing behavior need not be D-generated), intentions may fail (e.g., D-intended$_{gen,2}$ behavior need not be D-interpreted$_{gen,2}$), and so on. (c) My vocabulary can be expanded by combining in various ways the building blocks of generation, appearance, interpretation, and intention. For example, a behavior might be called 'D-faking' exactly if it is D-intended$_{gen,\infty}$ but not D-generated,[11] and a behavior might be called 'D-leaking' exactly if it is D-generated and D-appearing but neither D-intended$_{gen,0}$ nor D-intended$_{app,0}$.

---

[9] Under reasonable assumptions, saying that a behavior is D-intended$_{gen,\infty}$ exactly if it is D-intended$_{gen,i}$ for every $i \in \mathbf{N}$ is equivalent to saying that a behavior is D-intended$_{gen,\infty}$ exactly if it is intended to be D-interpreted$_{gen,\infty}$.

[10] Cf. Lewis (1978/1983: 266): "One act of storytelling might, however, be the telling of two different fictions: one a harmless fantasy told to the children and the censors, the other a subversive allegory simultaneously told to the *cognoscenti*."

[11] Equivalently: exactly if it is D-intended$_{gen,\infty}$ but not D-accompanied (a behavior being *D-accompanied* exactly if it is accompanied by D). One might object that the equivalence fails because D-accompanied behavior need not be D-generated; e.g., I may whistle *while* but *not because* I have D for you. I reply that in such cases the source's D is irrelevant to the behavior; this is not so when the behavior is D-intended$_{gen,\infty}$, so that a behavior which is both D-intended$_{gen,\infty}$ and D-accompanied is also D-generated.

2.1.3. The insulting intention

At the end of §2.1.1 I asked: what kind of intention makes a behavior insulting? Given the vocabulary I introduced in §2.1.2, I am now in a position to formulate the answer I will defend: a behavior is insulting exactly if it is D-intended$_{gen,\infty}$.

To start with, it seems necessary that a behavior be D-intended$_{gen,0}$ (i.e., intended to be interpreted as being D-generated) if it is to be insulting. I will argue, however, that this condition is not sufficient. Suppose that, in order to make you indirectly realize that I despise you, I arrange for you to hear me muttering insults about you in such a way that you believe you are *over*hearing me (i.e., you take me to believe that you are not hearing me). (E.g., we are alone in adjacent rooms with unusual acoustic properties of which I am aware but I know you are not.) My behavior is D-intended$_{gen,0}$, but is it insulting? It seems not, because it does not have the nature of a direct provocation that would leave you no choice but to react if you were to avoid losing face: even if my intention (that you interpret my behavior as being D-generated) succeeds, you may choose to pretend that you didn't hear me without losing face, since you take me to believe that you didn't hear me.

(Readers familiar with attempts to analyze the concepts of meaning and of communication may have recognized a similarity between my claim that a successful insulting intention must have the nature of a *direct* provocation and the claim that "true communication must be characterised as *wholly overt*": "either your behavior makes it clear that you are communicating, or else you are not truly communicating at all" (Sperber & Wilson 1995: 30). In fact, the discussion that follows, which essentially aims at explicating the relevant notion of directness, parallels the debate which has aimed at explicating the relevant notion of overtness.)

The above example indicates that, if a behavior is to be insulting, then the absence of the (second-order) intention that the behavior be interpreted as *not* being D-intended$_{gen,0}$ (i.e., as *not* being intended to be interpreted as being D-generated) is neces-

sary. The absence of this intention is not sufficient, however (contrast Grice 1969/1989a: 99, 1982/1989c: 303; Bach & Harnish 1987: 712). To modify the above example, suppose that I mutter my insults without caring whether you recognize my intention that you hear me. (I know that you don't recognize it but I don't intend that you fail to recognize it: I am accidentally, rather than as a result of premeditation, in a situation in which you are hearing me but you believe that you are overhearing me.) It seems again that my behavior is not insulting because it lacks the appropriate kind of directness.

Maybe what is needed (in addition to being D-intended$_{gen,0}$) to make a behavior insulting is not (just) the *absence* of the second-order intention that the behavior be interpreted as *not* being D-intended$_{gen,0}$, but the *presence* of the second-order intention that the behavior be interpreted as *being* D-intended$_{gen,0}$. In other words, maybe being D-intended$_{gen,1}$ suffices to make a behavior insulting. This proposal also fails, however (cf.: Strawson 1964/1971: 28-9; Grice 1969/1989a: 94-5; Sperber & Wilson 1995: 30). Suppose that, in order to make you believe that I plan to make you indirectly realize that I despise you, I pretend to misplace, in such a way that you will find and read it, a piece of paper which details a plan to make you hear me muttering insults about you in such a way that you will believe you will be overhearing me. If I go on to mutter these insults so that my behavior is D-intended$_{gen,1}$, my behavior is still not insulting: even if my scheme succeeds, you may choose to pretend that you didn't hear me without losing face, since you take me to believe that, although you heard me, you don't know that I intended you to hear me.

Similar counterexamples can presumably be produced if one adds the requirement that a *third*-order intention be present (i.e., if one suggests that being D-intended$_{gen,2}$ suffices for being insulting),[12] and so on ad infinitum[17] (cf.: Grice

---

[12] Rather than adding this requirement, one could require (following Grice 1957/1989b: 218-9) that the success of the second-order intention be intended to play a role in the success of the first-order intention. In other words, one could require the presence of the intention that the behavior be interpreted as being D-

1969/1989a: 95-9; Schiffer 1972: chap. 2). The response that I favor is to go all the way: as I said, I propose that a behavior is insulting exactly if it is D-intended$_{gen,\infty}$. One might object that this proposal has "little psychological plausibility. From the psychological point of view, intentions are mental representations capable of being realised in the form of action. No psychologist would want to analyse an utterance [or an instance of insulting behavior] as the realisation of an infinity of intentions" (Sperber & Wilson 1995: 31). I reply: why not? Take an analogy with beliefs. If you learn, e.g., that every integer whose decimal representation ends in 0 is divisible by 5, then you acquire at the same time infinitely many beliefs: for every integer *n* whose decimal representation ends in 0, you believe that *n* is divisible by 5. Saying that you have all these (infinitely many) beliefs is compatible with saying that you never entertain some of these beliefs,[13] and even with saying that you *cannot* entertain some of these beliefs (because, e.g., the decimal representations of some integers are too long and complex). A ground for ascribing to you beliefs that you do not and that you even cannot entertain is that you would agree that you have these beliefs were you (able) to entertain them.[14] Similarly for intentions, I submit. Resistance to the proposal that insulting behavior involves *having* infinitely many intentions may stem from misinterpreting this proposal as stating that insulting behavior involves *entertaining* infinitely many intentions.[15] A ground for ascribing to you

---

generated (partly) by means of being interpreted as being D-intended$_{gen,0}$. Adding this requirement does not suffice to make a behavior insulting, however. To modify the last example, suppose that rather than muttering insults I mutter phrases fraught with innuendoes, so that your interpreting my utterances as being D-generated depends crucially on your being forewarned of my plan (by reading the piece of paper). It seems again that my behavior is not insulting because it lacks the appropriate kind of directness.

[13] Cf. my distinction between occurrent and dispositional beliefs (§I.1.1.2): one *entertains* occurrent beliefs but one *has* dispositional beliefs.

[14] The proposition that every integer whose decimal representation ends in 0 is divisible by 5 entails (e.g.) the proposition that 23057922487610 is divisible by 5. This entailment, however, is *not* my (whole) reason for claiming that you believe the latter proposition if you believe the former. (In general, I am not claiming that the set of one's beliefs is closed under entailment.) My reason is rather that you would assent to the latter proposition were you to entertain it; the entailment might be (part of) your reason for assenting.

[15] Compare my distinction between the infinite chain of intentions one *has* and the finite chain of intentions

intentions that you do not and that you even cannot entertain is that you would agree that you have them were you (able) to entertain them.[16] The fact (if it is a fact) that beyond a certain level the intentions become too complex for you to entertain is no more an obstacle to ascribing them to you than the fact (if it is a fact) that the decimal representations of certain integers are too long and complex is an obstacle to ascribing beliefs about these integers to you.[17]

Another possible source of resistance to my proposal that insulting behavior involves an infinite series of intentions is the consideration that intentions should be "capable of being realised in the form of actions": how can one perform infinitely many actions? I reply that the same action may satisfy different intentions: it is sometimes possible to kill two (or more) birds with one stone. Shouting an obscenity to your face may satisfy every intention in the infinite series.

This completes my defense of the claim that a behavior is insulting exactly if it is D-intended$_{gen,\infty}$. So now we have a fairly good idea of what HEN prescribes: avoiding behavior likely to be interpreted as D-intended$_{gen,\infty}$.

---

one *entertains* to Lewis's (1969: 53-7) distinction between an infinite chain of reasons to form expectations and a finite chain of actually formed expectations in cases of coordination by agreement.

[16] I am *not* claiming that a ground for ascribing to you intentions that you do not and that you even cannot entertain is that you would agree that you have them were you presented with counterexamples like those I adduced above in the text. I have two reasons for avoiding this claim. First, accepting this claim might open the door to the implausible claim that the set of one's beliefs is closed under entailment (cf. footnote 14). Second, the counterexamples establish only that the *absence* of certain intentions (e.g., the second-order intention that the behavior be interpreted as *not* being D-intended$_{gen,0}$) is necessary (and that the presence of certain intentions is *not sufficient*) if a behavior is to be insulting.

[17] Grice (1969/1989a: 98-9) uses the claim that beyond a certain level some intentions become too complex for you to entertain in order to argue against the existence of an infinite series of counterexamples. Grice claims that in general "one cannot have intentions to achieve results which one sees no chance of achieving" (1969/1989a: 98). I see no chance of making you actually *interpret* my behavior as being D-intended$_{gen,i}$ for large enough *i*, so I cannot have the intention that you so interpret my behavior (or so Grice would say). I reply that making you "interpret" my behavior in a certain way is understood as making you *have*, rather than as making you *entertain*, a belief about my behavior; thus my intention that you interpret my behavior as being D-intended$_{gen,i}$ may have a good chance of being successful.

2.2. In what sense is HEN rationally warranted?

2.2.1. The tree framework and the problem of coarse-graining

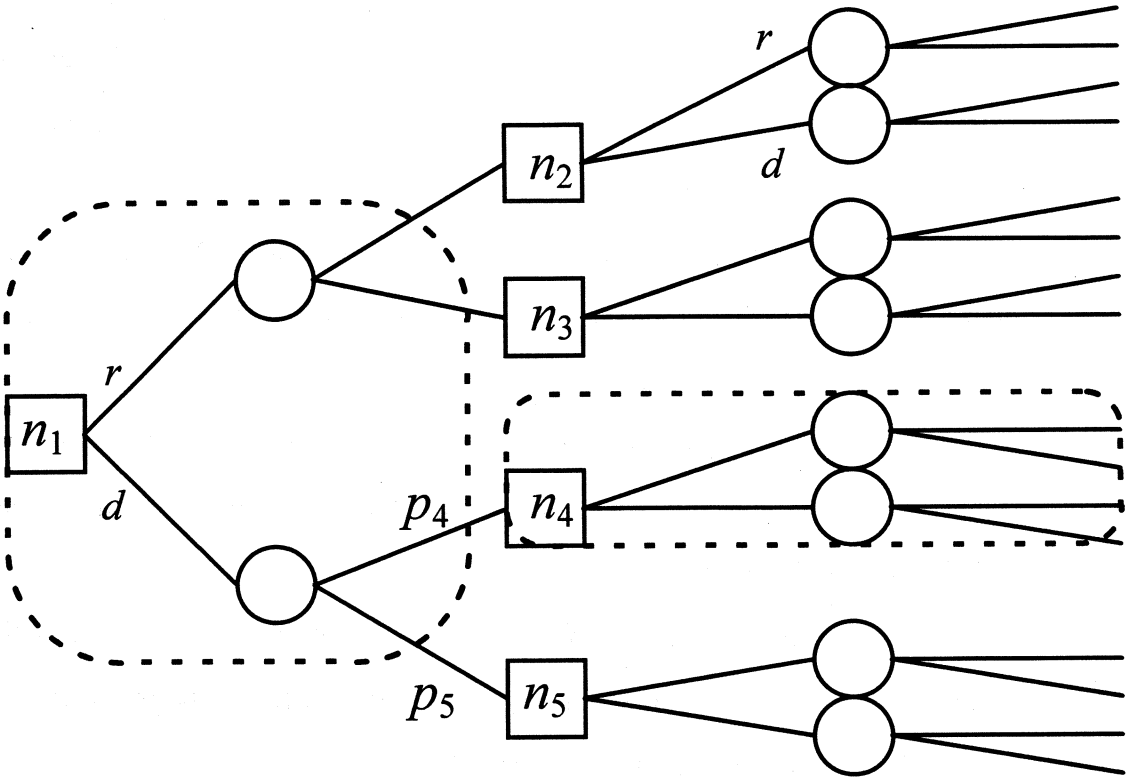


Figure 3.1. The tree framework

Figure 3.1 shows the framework I will be using, the "tree framework". Suppose that at time t_1 I am at decision node n_1. I can choose between various actions, some of which are disrespectful, *d*, and some of which are nondisrespectful, *r*. Then some event will happen whose outcome is not within my control; for example, others will retaliate or not if I insult them. So if I choose *d* then at the next time, t_2, with probability p_4 I will be at node n_4 and with probability p_5 I will be at node n_5. And so on for later times. Now a strategy tells me what to do at each decision node. For example, HEN tells me to choose an *r* branch at each decision node. In general there may be more than one *r* branch at a decision node, because there may be more than one way of behaving nondisrespectfully.

In such a case HEN does not tell me *which* of the *r* branches to choose; this is more formally the point that I made (in §1) when I said that this strategy is *coarse-grained.* A *fine*-grained strategy, by contrast, specifies a *unique* branch at each decision node. Expected utilities can be computed only for fine-grained strategies: a coarse-grained strategy corresponds in general to many different fine-grained strategies with different expected utilities. Returning to my example in §1, one can follow the coarse-grained strategy of not smoking and still lead either a very healthy or a very unhealthy life. So how can we evaluate a coarse-grained strategy, given that we can assign no unique expected utility to such a strategy? This is what I call "the problem of coarse-graining". I implicitly proposed in §1 a solution to this problem: I said that avoiding smoking is justified in the sense that the *healthiest* fully specified ways of living will not include smoking. More formally, the proposal is to evaluate a coarse-grained strategy by the expected utility of the *best* fine-grained strategies which are compatible with the coarse-grained strategy. I will now defend this proposal.

2.2.2. Two kinds of reason claims: bidirectional and unidirectional

I define a *reason sentence* as a sentence having the following (or some closely related) form: '*k* has (pro tanto) prudential reason to do/avoid *b*'. I define a *reason claim* as a proposition that can be standardly expressed by a reason sentence. I will argue that reason sentences are ambiguous: for every reason sentence there are two reason claims that the sentence can standardly express.

Before I explain the ambiguity, some background is needed. Consider an agent and a context in which some (short- or long-term) alternatives are open to the agent. For example, if a student requests a recommendation letter from me, I can refuse or accede to this request. Some alternatives can be "decomposed" into more specific ones: a recommendation letter can be lukewarm or enthusiastic, an enthusiastic letter can contain only positive statements or also some negative ones, and so on. Given that I must perform

some specific possible action, the "decomposition" must stop at some maximally specific alternatives[18] (m.s.a.'s);[19] I conceptualize an alternative in general as a union of m.s.a.'s.[20] Now a (pro tanto) prudential reason R to maximize a welfare function W[21] can be associated with an *optimality set*, O_R, consisting of all and only those m.s.a.'s which maximize W. The set O_R partitions the union of all m.s.a.'s into the union of those m.s.a.'s that "satisfy" R and the union of those m.s.a.'s that "violate" R. Note that an alternative which is not maximally specific may neither satisfy nor violate R, in the sense of being neither a subset of O_R nor a subset of the complement of O_R.

Given the above background, here are the two reason claims that can be standardly expressed by the reason sentence (T) 'k has (pro tanto) prudential reason to do b': (1) there is a reason (which applies to k and)[22] whose optimality set is b (*bidirectional* reason claim; formally: $\exists D\ b{=}O_R$); (2) there is a reason whose optimality set is included in b (*unidirectional* reason claim; formally: $\exists D\ b{\supseteq}O_R$).[23] For example, take b to be the

---

[18] One might object: "If it is my duty to pay you ten dollars then I have latitude in that I may pay by cash, check, or money order; or if it is my duty to pay you in cash, then I may pay by giving you a ten, or fives, or ones; or if it is my duty to give you a ten, then I may give you this one, that one, or the other one; or if it is my duty to give you this one, then I may hand it to you with the face looking up, or down, or right, or left; and so on, *ad infinitum*" (Chisholm 1963: 4, quoted in Hill 1971: 63). But even if there are infinitely many m.s.a.'s, and even if some m.s.a.'s cannot be *described* in a finite (or in a unique) way, it remains the case that no m.s.a. can be decomposed: if I give you a ten-dollar bill, I give you a specific bill under specific circumstances and in a specific way.

[19] I understand m.s.a.'s as *sets*. What are their members? It doesn't matter for present purposes; it matters only that the m.s.a.'s be in one-to-one correspondence with the specific possible actions. (E.g., the m.s.a.'s can be singletons whose members are the specific possible actions themselves.)

[20] I am not taking a stand on whether *every* union of m.s.a.'s is an alternative; *maybe* some miscellanies of m.s.a.'s should not be considered alternatives because they could not be used in formulating guidelines on behavior.

[21] The reason is pro tanto but not conclusive if W takes into account some but not all factors contributing to the agent's welfare.

[22] For the sake of simplicity, I omit the parenthesized clause in the sequel.

[23] Note that: the bidirectional reason claim entails the unidirectional one but not vice versa; a reason claim cannot be both bidirectional and unidirectional. I am *not* saying that there are two kinds of *reasons* (as op-

alternative of not smoking. By uttering T one standardly claims that there is a reason R which is violated by (any m.s.a. included in) smoking. But in the bidirectional use of T one further claims that R is satisfied by *all* m.s.a.'s included in not smoking; whereas in the unidirectional use of T one leaves open the possibility that R is also violated by some m.s.a.'s included in not smoking. This example supports the conclusion that the more natural and interesting understanding of reason sentences is the unidirectional one. I will now adduce a further consideration to support this conclusion.

The following sentence seems to express a plausible principle: *(P) Necessarily, if k has reason to do b′ and k cannot do b′ without doing b, then k has reason to do b.* For example: if I have reason to meet you and I cannot meet you without catching the noon train, then I have reason to catch the noon train. Now it can be shown that:

Theorem 3.1. P is true (i.e., expresses a true proposition) exactly if in P 'k has reason to do b' is used unidirectionally.[24]

Given Theorem 3.1, if one accepts that P is true then one should accept that at least some reason sentences are used unidirectionally. Intuitively: if I assert that I have reason to not smoke *only because* I believe that I have reason to act so as to be healthy, then my assertion presupposes the existence of a reason which I can violate even without smoking (e.g., by taking drugs), so that (the reason claim which I express by) my assertion is unidirectional. On the other hand, if my assertion that I have reason to act so as to be healthy is not derived from any other belief about my reasons, then it is bidirectional.

---

posed to *reason claims*): I take every reason to perform the same job, which is partly to partition the union of all m.s.a.'s into those that satisfy and those that violate the reason.

[24] *Proof.* Let T = 'k has reason to do b' and T' = 'k has reason to do b''. (1) If T is used unidirectionally, then: (i) if T' is used unidirectionally, then P is true: $\Box[(\exists R\ b' \supseteq O_R)(b \supseteq b') \rightarrow (\exists R\ b \supseteq O_R)]$; (ii) if T is used bidirectionally, then P is again true: $\Box[(\exists R\ b' = O_R)(b \supseteq b') \rightarrow (\exists R\ b \supseteq O_R)]$. (2) If T is used bidirectionally, it can be similarly seen that P is false. (Formalizing 'k cannot do b' without doing b' as '$b \supseteq b'$' is compatible with using 'cannot' to express any usual kind of modality—e.g., logical, metaphysical, physical, or psychological. Strictly speaking, the proof needs the assumption that T and T' have no standard uses besides the unidirectional and the bidirectional ones.) □

This example suggests a criterion for finding out whether any specific reason sentence is used bidirectionally or unidirectionally: roughly, underived reason sentences are used bidirectionally, whereas derived reason sentences are in general used unidirectionally. Given that few reason sentences are underived, reason sentences are normally understood unidirectionally.

If the above is right, then to say that I have reason to follow a coarse-grained strategy is normally to say that the *best* fine-grained strategies (long-term m.s.a.'s) are included in the coarse-grained strategy. This justifies my proposal to evaluate a coarse-grained strategy by the expected utility of the best fine-grained strategies which are compatible with the coarse-grained strategy. So now our object becomes to find the expected utility of the best fine-grained strategies which are compatible with the coarse-grained strategy of HEN and to compare this expected utility with the expected utility of the best fine-grained strategies which are compatible with each competing coarse-grained strategy.

3. HEN is better than HED

My argument for the superiority of HEN to HED has two premises, a mathematical and an empirical one. The mathematical premise is that, if a certain condition *K* (to be specified later on) holds, then HEN is better (i.e., has higher long-term expected utility) than HED. The empirical premise is that condition *K* holds. In §3.1 and §3.2 I defend these two premises in turn.

3.1. The mathematical premise

3.1.1. Long- and short-term expected utilities

A fine-grained strategy corresponds to a set of paths in the decision tree. The *long*-term expected utility of a fine-grained strategy is the weighted sum of the utilities of the paths that correspond to the strategy times the probabilities of these paths. What is

the utility of a path? Since we are talking long-term here, each path corresponds to a possible course of life, so the utility of a path is a measure of the quality of the corresponding course of life. Now each path is a concatenation of path-segments which correspond to the periods between successive times. In Figure 3.1, for example, one of the path-segments between time t_1 and time t_2 is the one from node n_1 to node n_2. The utility of a path-*segment* can be defined as a measure of the quality of one's life at the time which corresponds to the end of the segment. Now to each decision node corresponds a subtree which contains the path-segments that start from the given node. For example, the dotted lines in Figure 3.1 give the subtrees that correspond to nodes n_1 and n_4. The *short*-term expected utility of a fine-grained strategy at a given decision node can be defined as the weighted sum of the utilities of the path-segments in the subtree that correspond to the action prescribed by the given strategy at the given node times the probabilities of these path-segments. In Figure 3.1, for example, the short-term expected utility at node n_1 of the strategy of always choosing the lowermost branch is the weighted sum of the utility of the path-segment from n_1 to n_4 times its probability, p_4, plus the utility of the path segment from n_1 to n_5 times its probability, p_5. So now we have definitions of the long- and short-term expected utilities of a strategy.

Let the *Additivity Condition* be that the utility of a path is the sum of the utilities of its constitutive path-segments. In other words, measuring the quality of a whole life amounts to measuring the qualities of the life at various times and taking the sum. This condition has been criticized in the literature, but later on I will explain that my argument probably goes through even if the condition fails. For the moment let me assume that the condition holds.

Theorem 3.2. *If the Additivity Condition holds, then the long-term expected utility $EU_L(S)$ of a fine-grained strategy S is a weighted sum of the short-term expected utilities $EU_n(S)$ of the strategy at the decision nodes n of the tree, the weight at node n being the probability $P_S(n)$ that node n is reached if strategy S is followed:*

$$EU_L(S) = \Sigma_n P_n(S) EU_n(S).^{25}$$

The above theorem gives us a way of expressing the long-term expected utility of a strategy in terms of its short-term expected utilities at the nodes of the tree. This means that, under certain conditions, we can compute the *difference* between the *long*-term expected utilities of two strategies by computing the differences between the *short*-term expected utilities of these strategies; in other words, by comparing the strategies on a node-by-node basis.[26] I proceed now with such a node-by-node comparison.

3.1.2. Short-term comparison of HEN and HED

Consider the subtree which corresponds to a given node, say node n_1. To fix ideas, take a concrete example. Suppose that when I try to check in for my flight the airline agent tells me the flight is overbooked. I am trying to decide what to do. I can behave in various nondisrespectful ways; for example, I can explain politely to the agent that I *need* to get into this flight because I am expected to give a talk at a conference. I can also behave in various disrespectful ways; for example, I can start yelling at the agent. In Figure 3.2 for simplicity I show only one fine-grained nondisrespectful alternative, r, and only one fine-grained disrespectful alternative, d. If I choose r, then with probability p_2 I will get the result I want, a seat in the flight; call this outcome c, for *compliance*. With probability p_3, however, which is $1\text{-}p_2$, I will not get a seat; call this outcome z, for *resistance*. If on the other hand I choose d, then with probability p_4 the agent will give me a seat, although probably reluctantly, to avoid embarrassment, so call

---

[25] I omit the proof; the reader may want to check as an exercise that the equality asserted by Theorem 3.2 holds for the tree of Figure 3.1 and the strategy of always choosing the lowermost branch.

[26] In general different nodes are reachable by different strategies; e.g., in the tree of Figure 3.1, no node at the top of the figure is reachable by HED and no node at the bottom of the figure is reachable by HEN. So the only common node in the sums giving EU_L(HEN) and EU_L(HED) will be n_1. We can still compare the two strategies on something like a node-by-node basis, however, if some assumptions of symmetry etc. between the top and the bottom of the figure are made.

this outcome *s*, for *submission*. Finally, with probability p_5 the agent will not give me a seat if I yell and will even retaliate, maybe not directly, by yelling herself, but rather indirectly, by helping me less than she could. So call this outcome *t*, for *retaliation*.[27]

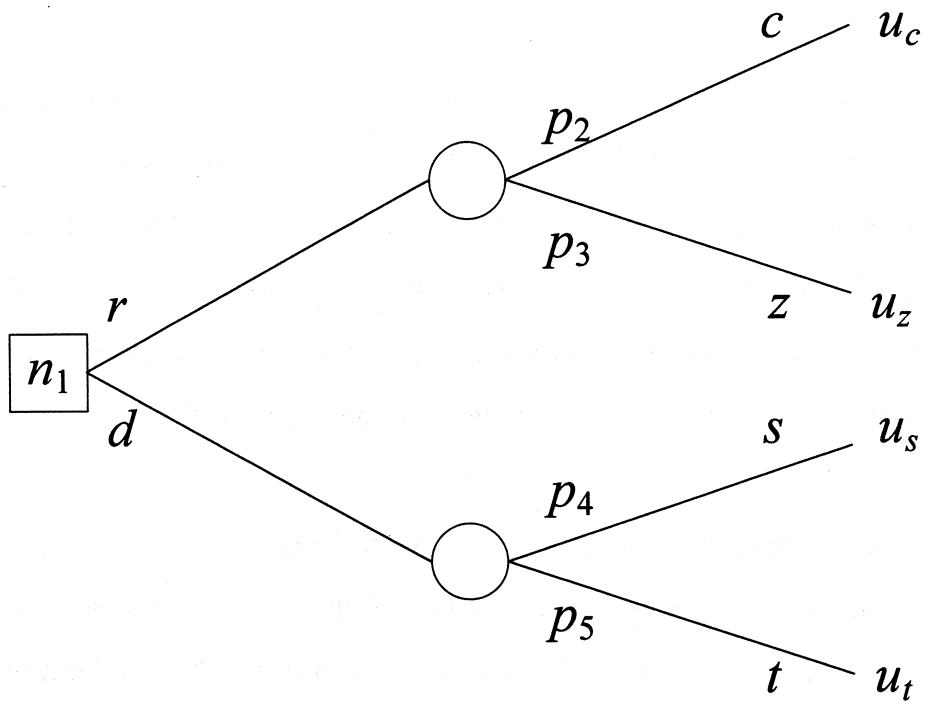


Figure 3.2. Short-term comparison of HEN and HED

The short-term expected utility of *r*, and thus of any strategy which prescribes *r* at node n_1, is p_2 times the utility of compliance, u_c, plus p_3 times the utility of resistance, u_z; similarly, the short-term expected utility of *d* is $p_4u_s+p_5u_t$. Note that the utility of submission is about the same as the utility of compliance: in both cases I get a seat in the flight. But the utility of retaliation is less than the utility of resistance: in retaliation the agent

[27] Strictly speaking I need an extra chance node with a certain probability that the agent will try to harm me *in addition* to not giving me a seat; but it turns out that this makes no difference, so for simplicity I omit this extra node.

may try to harm me *in addition* to not giving me a seat. Let u_T be the (positive) difference between the utilities of resistance and retaliation, u_z-u_t. Let me also denote p_2 by $p_{c|r}$, the probability of compliance given that I choose r, to make explicit that this probability may depend on r, on *which* nondisrespectful alternative I choose; similarly, denote p_4 by $p_{s|d}$. Then one can show that:

Theorem 3.3. $EU_{n1}(HEN)$-$EU_{n1}(HED)$=$EU_{n1}(r)$-$EU_{n1}(d)$=$(p_{c|r}$-$p_{s|d})(u_c$-$u_z)$+p_5u_T.[28]

The second term of the above sum is positive: it corresponds to the expected cost of retaliation. The first term is a product of two differences. The difference between the utilities of compliance and resistance is clearly positive, so the crucial question is whether the difference of probabilities $p_{c|r}$-$p_{s|d}$ is positive: is the agent more likely to give me a seat if I ask politely or if I yell? I will return to this question in §3.2.

3.1.3. Long-term comparison of HEN and HED

Call a node *nondisrespect-propitious* (NP) exactly if the short-term expected utility of a best nondisrespectful alternative at that node, r^*, is at least as high as the short-term expected utility of a best disrespectful alternative at that node, d^*. Call a node *disrespect-propitious* (DP) exactly if it is not NP. Now let us compute the difference between the *long*-term expected utilities of the coarse-grained strategies HEN and HED. Call R^* a best fine-grained strategy compatible with HEN and D^* a best fine-grained strategy compatible with HED. Theorem 3.1 gives the long-term expected utility of R^*—and thus also of HEN, given my solution to the problem of coarse-graining (§2.2.2)—as a sum of the short-term expected utilities of R^*; similarly for HED. By breaking down these sums into two parts, corresponding respectively to NP and DP nodes, and making some further assumptions, it can be shown that $EU_L(HEN)$-$EU_L(HED)$ is positive if,

---

[28] *Proof.* $EU_{n1}(r)$-$EU_{n1}(d)$ = $(p_2u_c$+$p_3u_z)$-$(p_4u_s$+$p_5u_t)$ = $(p_2u_c$-$p_4u_s)$+$(p_3u_z$-$p_5u_t)$ = $(p_2$-$p_4)u_c$ +$(1$-$p_2)u_z$-$(1$-$p_4)(u_z$-$u_T)$ = $(p_2$-$p_4)(u_c$-$u_z)$+$(1$-$p_4)u_T$, from which Theorem 3.3 immediately follows. □

first, (K_1) *most* nodes are NP and, second, (K_2) no DP node is a disrespect *sink*; in other words, for no DP node is the short-term expected utility of a best disrespectful alternative *much* higher than the short-term expected utility of a best nondisrespectful alternative. The conjunction of K_1 and K_2 is condition K, so we have reached the result that if condition K holds then HEN is better than HED. This, you may recall, is the mathematical premise. This premise is intuitively clear: if at most nodes disrespect is worse than nondisrespect and even at the few nodes at which disrespect is better it is not *much* better, then the strategy of behaving disrespectfully should be worse than the strategy of behaving nondisrespectfully.

Let me now return to the Additivity Condition (see Theorem 3.2). I said in §3.1.1 that this condition has been criticized in the literature. Velleman (1991/2000: 58-9), for example, argues that this condition is false because a life which improves over time is better than a life which deteriorates over time even if the sums over time of measures of the instantaneous qualities of the two lives are the same. Let me grant this for the sake of argument. My reasoning still probably goes through, however. First, because even if the first life is better than the second, it seems that it will not be *much* better, so only a small correction is needed. Second, because I see no reason to suppose that this small correction will favor HED over HEN.

3.2. The empirical premise

3.2.1. Most nodes are nondisrespect-propitious

Let me start with the first of the two parts of condition K, namely the claim that most nodes are NP. From my earlier node-by-node comparison of strategies (§3.1.2, Theorem 3.3) it follows that a *sufficient* condition for a node to be NP is that the following inequality holds: $p_{c|r^*} > p_{s|d^*}$. In words: a best nondisrespectful alternative is more likely to result in compliance than a best disrespectful alternative is likely to result in

submission. Recall that in the overbooked flight example the corresponding question was: am I more likely to get a seat if I ask politely or if I yell?

A first reaction of many people to the above question is that *of course* many such situations are *dis*respect-propitious: yelling often gets results. But although it is uncontroversial that yelling often gets results, this does not settle the relevant *comparative* question of whether some nondisrespectful alternatives are *more* likely than yelling to get results. In response one might point out that in some cases we try to get results respectfully, we fail, and then we resort to disrespect and we succeed. For example, some people may have had the experience that *only* after resorting to yelling did they manage to get a seat in an overbooked flight. But again this does not settle the relevant question: maybe the nondisrespectful way in which they were trying to get a seat before they resorted to yelling was not a *best* nondisrespectful way. Of course *I* have the burden of defending the claim that a best nondisrespectful way is better than a best disrespectful way; but for the moment I am making the preliminary point that contrary to appearances it is hard to find clear cases where my claim is false, clear DP nodes.

A second and very common reaction to my question of whether asking politely or yelling is more likely to get me a seat is that it depends. It depends, for example, on the personality of the agent with whom I am dealing: some agents will be much more likely than others to be intimidated if I yell. True, but the relevant question is *what I can reasonably expect*, not *what is in fact the case* about the agent: it is my *subjective* probabilities which determine what I should do (cf. Gibbard 1971/1990a: 28-42, 1990b: 42-3). And these subjective probabilities depend, first, on what I know about the base rates (namely the percentages of people in general who would react in this or that way if I yelled to them), and second on whether I have specific information about the agent in front of me which should make my probabilities deviate from the base rates. One might argue that in many cases I do have such specific information: the agent in front of me may look to me like a submissive person, or I may even have interacted with her repeat-

edly in the past and she has always submitted when I yelled. I reply again that this does not settle the relevant *comparative* question, because typically I don't know whether she *would* have complied if I had behaved respectfully. Moreover, there is extensive psychological evidence that, although we strongly believe otherwise, we are poor predictors of people's behavior in *unobserved* situations (see §II.4.3). The upshot of this discussion is that the most reasonable thing to do is to set my subjective probabilities approximately equal to the base rates. So we need empirical evidence addressing the question: how do people in general react to respect and to disrespect? I will now examine such empirical evidence.

Stimulus↓ \ Response→	HD	HS	FS	FD
HD: hostile-dominant	**5**	3	1	1
HS: hostile-submissive	1	2	0	**5**
FS: friendly-submissive	1	1	4	**6**
FD: friendly-dominant	0	2	**5**	2

Table 3.1. Empirical evidence on base rates

Orford (1986) reviewed ten studies[29] examining various kinds of interactions between people: interactions between therapists and clients, interactions within family dyads consisting of a father and a mother, a father and a son, or a mother and a son, interactions within maritally satisfied and dissatisfied couples, within groups of acquainted female students, and interactions of hyperaggressive boys both with other such boys and with staff during residential treatment. Almost all of these studies were naturalistic, in

[29] Namely: Billings 1979; Blumberg & Hokanson 1983; Crowder 1972; MacKenzie 1968; Mueller 1969; Mueller & Dilling 1968; Raush, Dittman, & Taylor 1959; Rice 1969; Shannon & Guerney 1973; Swenson 1967.

the sense that the participants did not know that a study was taking place. The observed behaviors were grouped into four categories: hostile-dominant, hostile-submissive, friendly-submissive, and friendly-dominant. Table 3.1 gives the results, in terms of the number of studies in which the corresponding combination of behaviors was the most frequent or most significant. Take first hostile-dominant behavior, which would correspond in my example to trying to get a seat by yelling. Such behavior did often result in submission: yelling does often get results. But the *most* frequent response to hostile-dominant behavior was not submission: it was rather *further* hostile-dominant behavior. Take now friendly-dominant behavior, which would correspond in my example to trying to get a seat respectfully but firmly. The most frequent response to such behavior was in fact friendly submission. Note that a friendly-*submissive* stimulus was most often met with a friendly-*dominant* response: protesting too meekly is relatively unlikely to get results. This may give the impression that nondisrespect is less effective than disrespect. But this, as I said, is the wrong comparison: we must compare hostile-dominant with friendly-*dominant*, not with friendly-*submissive* behavior. And the results of the *right* comparison support my case. If the numbers in the table can be used to compute very rough estimates of conditional probabilities, then the probability of a submissive response given a *friendly*-dominant stimulus comes out 7/9, much higher than the probability of submissive response given a *hostile*-dominant stimulus, which is only 4/10; this is the inequality we want for nondisrespect-propitiousness. Let me make clear that I am not proposing these results as conclusive: for example, friendly and hostile behavior do not correspond exactly to nondisrespect and disrespect. But I do think that the results shift the burden of proof to opponents of the claim that most nodes are NP.

Recall that nondisrespect-propitiousness amounts to the positivity of the following sum: $(p_{c|r^*}-p_{s|d^*})(u_c-u_z)+p_5u_T$. I argued so far that the difference of probabilities $p_{c|r^*}-p_{s|d^*}$ is positive; this is sufficient for nondisrespect-propitiousness because then both terms of the above sum are positive. But we can get nondisrespect-propitiousness even if

the difference of probabilities and thus the first term of the sum is negative, provided that the *second* term of the sum, namely the expected cost of retaliation, is high. I will now adduce some empirical evidence suggesting that the expected cost of retaliation is indeed high. According to Daly and Wilson, "the most prevalent variety of urban homicide in the United States" is "altercation of relatively trivial origin: insult, curse, jostling, etc." (1988: 125). Of course I am not saying that if you insult the airline agent you can expect her to kill you, but these data on homicides suggest that insulting people is not without risk. Similarly, Baumeister, Smart, and Boden (1996; cf. Baumeister 2001; Bushman & Baumeister 1998), in an interdisciplinary review of the evidence, have argued that aggression, crime, and violence are commonly a result of "threatened egotism", namely highly favorable views of the self that are disputed by some person or circumstance, for example by means of an insult or a negative comment. It is important to note that retaliation can take indirect forms when the aggressor is too powerful. Greenberg (1990), for example, found in an experimental study that employee theft increased much more in reaction to an arbitrary than to an adequately explained pay cut. Taken together, these results provide some indirect support for the claim that the expected cost of retaliation is high and thus provide further support for the claim that most nodes are NP.

3.2.2. There are no disrespect sinks

Having completed my defense of the first component of condition *K*, I turn now to the second component: the claim that there are no disrespect "sinks". Recall that a node is a disrespect sink exactly if a best disrespectful alternative at that node is *much* better than a best nondisrespectful alternative. One way in which this can happen is if the best nondisrespectful alternatives have very low expected utility; for example, you survive only by being disrespectful. One might argue that this happens in "macho" environments, in which aggression is very widespread. In such environments you *need* to behave disrespectfully in order to establish a deterrent reputation, to be recognized as a per-

son not to be messed with; nondisrespect is taken as a sign of weakness and you become an easy target. I think, however, that this argument is flawed. Avoiding disrespect does not amount to meekly swallowing insults; I can establish a deterrent reputation without behaving disrespectfully. How? By never initiating insults myself, and by always responding to an insult with a noninsulting warning that I am ready to fight if my opponent persists—and then (and this is the crucial point) consistently making good this warning whenever my opponent does persist. It is true that I can *also* establish a deterrent reputation *dis*respectfully, by preventively insulting people and picking up fights. But the nondisrespectful strategy is in the long run *better* than the disrespectful one, because it is associated with a lower probability of retaliation: whenever I am insulted *after* it has become known that I am a good fighter, a fight is less likely to ensue if I respond with a warning than if I respond with an insult: a warning is less likely than an insult to make my opponent feel compelled to continue the altercation in order to save face. One might respond that if I don't project an image of toughness then I may be dead *before* I have the time to establish a deterrent reputation noninsultingly. I don't think so, however. Compare two newcomers to the macho environment who are equally good fighters but don't yet have any reputation. One of them sits quietly in a corner, while the other goes around insulting people. It seems that the latter is more likely than the former to get involved into fights and thus to be killed, given that they are equally good fighters.

This completes my defense of the empirical premise, and thus also of the superiority of HEN to HED.

4. Extensions of the argument

In this final section I argue very briefly that HEN is better than three other strategies I mentioned in §1: CBC, IDH, and TFT.

(1) Why commit ourselves in advance to always behaving in the same way, namely nondisrespectfully, as HEN prescribes? Why not adopt instead the more flexible

case-by-case (CBC) strategy of deciding how to behave in each case anew? The difference in flexibility between CBC and HEN is less than meets the eye: like CBC, HEN prescribes deciding how to behave in each case anew—although subject to the restriction of considering only nondisrespectful alternatives. But there is a more important reason why CBC seems better than HEN: it seems that adopting CBC will make us (a) behave disrespectfully (and thus outperform HEN) at DP nodes but (b) behave nondisrespectfully (and thus equiperform HEN) at NP nodes. In reply I contest (b): there are two reasons why adopting CBC will often make us behave disrespectfully even at NP nodes. First, disrespect is "addictive": strategically insulting at DP nodes will often make us lose our temper and insult even at NP nodes. The point is simply that with disrespect, as with certain drugs, selective abstinence is much harder than total abstinence. Second, effective disrespectful alternatives are frequently more salient (easier to think of) than effective nondisrespectful ones: witness the initial appeal of HED in macho environments (§3.2.2). Adopting HEN will make us look hard for effective nondisrespectful alternatives, so we will eventually become quite skilled at finding them (practice makes "perfect"). But adopting CBC will prevent us from developing this skill to the same extent: we will often settle on the salient effective disrespectful alternatives, mistakenly believing that the node at hand is DP. So (given also the absence of disrespect sinks) adopting CBC will in practice make us gravitate towards HED and thus on average do worse than with HEN.

(2) Rather than deciding in each case anew whether to behave disrespectfully or not, we might identify a *reliable indicator*—a strong positive correlate—of disrespect-propitiousness and decide in advance to habitually behave disrespectfully whenever this indicator is present and nondisrespectfully whenever it is absent. It is possible to form such compartmentalized habits: an executive may habitually treat subordinates condescendingly and superiors deferentially. If the presence and the absence of the indicator are readily detectable, then such a strategy of *indicator-differentiated habits* (IDH) can

outperform HEN: IDH largely escapes the two problems that I raised against CBC. The first problem was that adopting CBC will often make us lose our temper and behave disrespectfully against our better judgment. This problem was partly caused by CBC's prescription to always keep disrespect a live (hence often tantalizing) option until deciding to behave nondisrespectfully; but with IDH disrespect is not an option when we detect the absence of the indicator. The second problem was that adopting CBC will often make us settle on salient efficient disrespectful alternatives even at NP nodes; but with IDH our search is *confined* to nondisrespectful alternatives when we detect the absence of the indicator. Now what could be a reliable indicator of disrespect-propitiousness? A plausible candidate is the *submissiveness* of our interaction partner: if we are more likely to achieve our goals by treating her disrespectfully than by treating her nondisrespectfully, then the node at hand is probably DP (and vice versa). But is submissiveness readily detectable? One might think so on the ground that people are *translucent* (cf. Gauthier 1986: 174): we can usually detect, though not with certainty, how they are disposed to behave towards us. But I deny that the kind of translucency which people possess makes submissiveness readily detectable. You may be translucent in the sense that your meek demeanor enables me to accurately infer that I am likely to achieve my goals if I treat you disrespectfully. It doesn't follow that you are submissive, because I may be even *more* likely to achieve my goals if I treat you *non*disrespectfully. Such *comparative* judgments of likelihood are usually beyond our epistemic reach because (as I argued in §II.4.3) we are poor predictors of people's behaviors in *unobserved* situations. So it seems that no readily detectable indicator is forthcoming; IDH looks like a great but impractical idea.

(3) An interaction can consist of multiple moves. I mock you (move 1); you curse me (move 2); I slap you (move 3); you shoot me (move 4); end of interaction. A *microstrategy* prescribes how to behave at each move of an interaction; a *macrostrategy* prescribes how to behave at each interaction; so a macrostrategy prescribes microstrate-

gies. The four macrostrategies I have examined so far (HEN, HED, CBC, and IDH) were implicitly understood as prescribing *homogeneous* microstrategies: behave in the same way at each move of an interaction (e.g., for HEN, in an effective nondisrespectful way). Another foil for HEN, the macrostrategy of *tit for tat* (TFT), prescribes a *heterogeneous* microstrategy: behave nondisrespectfully at the first move (if the first move is yours), and at each of your other moves behave disrespectfully exactly if your interaction partner has behaved disrespectfully at the previous move. TFT differs from Axelrod's (1984) TIT FOR TAT: Axelrod talks about alternatives like cooperation and defection, not about nondisrespect and disrespect. Unlike TFT, TIT FOR TAT is in a sense compatible with HEN: maybe we should always reciprocate defection *nondisrespectfully*. TFT—like TIT FOR TAT—has four apparently attractive features. Three of them are shared by HEN: TFT is *simple* (so apparently we create clear expectations), *nice* (we never *initiate* disrespect, so we avoid unnecessary trouble), and relatively *forgiving* (we always reciprocate a conciliatory nondisrespectful move, so apparently we escape retaliation spirals). But a fourth feature of TFT makes all the difference: unlike HEN, TFT is *retaliatory* (we always reciprocate disrespect, so we deflect attempts to exploit us). Not only does this feature confer on TFT no advantage relative to HEN (*effective* nondisrespect also deflects attempts to exploit us), it also proves the demise of TFT: reciprocating disrespect is likely to cause retaliation spirals. This is because of what Fisher and Brown (1988: 25-30) call "partisan perceptions" or what Baumeister (1997: 18-9) calls "the magnitude gap": offenses seem more important to victims than to perpetrators. I perceive your insult as harsher than you think it is; I think my reciprocating insult is just as harsh as your original insult was and is thus justified; you perceive my reciprocating insult as harsher than I think it is, hence as harsher than you think your original insult was, hence as unjustified; and the escalation ball keeps rolling. TFT's forgiveness is of little use: you are unlikely to make a conciliatory move and thus to enable me to exercise this forgiveness. TFT's simplicity is also of little use: even if you realize that I think my reciprocating in-

sult was justified, you will probably still be outraged by the "fact" that it was excessive. So HEN is better than TFT.

I have not examined all possible competitors to HEN, but I think that the above considerations can provide an inductively strong argument for the conclusion that HEN is better than every competing strategy and thus for my thesis that we have strong prudential reason to follow HEN—to avoid disrespect.

APPENDIX
A TRIPARTITE DEFINITION OF ATTITUDES

Tripartite definitions of attitudes are well entrenched in the social psychological literature (Greenwald 1989a: 6; Zanna & Rempel 1988: 316). Recently, however, such definitions (or models)[1] of attitudes have come under attack, to the point that Greenwald (1989a: 6) claims that "a harsh evaluation of the three-component definition may be warranted" and Tesser and Shaffer (1990: 480) speak of "trashing the tripartite definition of attitudes". More specifically, it has been suggested that tripartite definitions of attitudes (1) promote confusion about the attitude-behavior relationship (Greenwald 1989a, 1989b; Zanna & Rempel 1988), (2) are unclear (McGuire 1989; Pratkanis 1989) and unparsimonious (Cacioppo, Petty, & Geen 1989), and (3) may conflict with the results of empirical studies (Cacioppo et al. 1989; Dillon & Kumar 1985; Eagly & Chaiken 1993; McGuire 1989; Mendler, Doll, & Orth 1990). In this appendix I defend the following tripartite definition of attitudes against the above three kinds of attack:

> An attitude in general consists of three components: an affective, a cognitive, and a conative one. These components are dispositions to have certain affective reactions, beliefs, and motives respectively.

My definition is able to meet the attacks partly because it differs from traditional tripartite definitions in three main respects. First, I define the conative component as a disposition to have certain *motives*, not as a disposition to *behave* in certain ways. Second, I don't presuppose that a common evaluative disposition underlies the three

---

[1] Some social psychologists speak of tripartite (or "three-component") *models*, rather than *definitions*. It seems that these two terms are frequently used interchangeably (e.g., Zanna & Rempel 1988: 316).

components. Third, I claim that an attitude has three components *in general*, not *always*. The relevance of these three points should become clear in the sequel.

In §1, §2, and §3 I address respectively the above three kinds of attack.

1. Objections concerning the attitude-behavior relationship

The claim that tripartite definitions of attitudes promote confusion about the attitude-behavior relationship is expressed powerfully by Greenwald (1989a: 6):

> Consider that the following four types of operations involving action in relation to an attitude object can serve equally *either* to measure the conative (behavioral) component of an attitude *or* to measure behavior that is presumably under the control of that attitude component: (a) observations of overt action, (b) verbal self-report of past action, (c) self-report of intentions regarding action, and (d) endorsement of statements about hypothetical actions. With this range of operations, a single research investigation can serve to test (a) the attitude-behavior relationship, (b) relations of the conative to other attitude components, or (c) the relation between behavior and the conative component of attitude. By affording this multiplicity of interpretations, the three-component definition appears to permit too broad an array of interpretations for a given set of data.

Greenwald's criticisms fail if one understands (as I do) the conative component as a disposition to have certain *motives*. A motive can be outweighed by opposing motives and thus need result neither in the formation of a corresponding intention nor in the performance of a corresponding behavior. To modify one of Greenwald's (1989a: 7-8) own examples, a motive to quit my job resulting from an attitude of dislike toward my job may never lead me to quit my job if it's outweighed by a motive to support my family by the income that I get from my job. Now it should be clear that, on my understanding of the conative component, Greenwald's "four types of operations" *don't* serve equally well to measure the conative component and behavior: (a) observations of overt action and (b) verbal self-reports of past action measure behavior *rather than* the conative component. (c) Self-reports of intentions are trickier. Sometimes social psychologists lump intentions together with behavior: "Behavioral responses also can be

regarded as encompassing *intentions* to act that are not necessarily expressed in overt behavior. For example, an individual may intend to circulate a petition tomorrow, but may or may not actually carry out this intention" (Eagly & Chaiken 1993: 12). To my mind, this practice of lumping intentions together with behavior is problematic: intentions are intermediate links in causal chains that may lead from motivation to behavior, so that measuring intentions does not enable one to distinguish motivation from behavior, let alone study the relationship between the two. Concerning finally (d) endorsements of statements about hypothetical actions, they seem again to measure intentions rather than behavior, and they are thus not directly relevant to empirical studies of the attitude-behavior relationship. I am not suggesting that measures of intentions be discarded in psychological studies: they could be used to investigate the attitude-*intention* relationship, which might be of interest in its own right. It's the failure to distinguish between motives, intentions, and behavior, not my tripartite definition of attitudes, that promotes confusion about the attitude-behavior relationship.

Greenwald might object that motives are elusive: how are we to measure them? I reply that motives are no more elusive than intentions: if Greenwald is happy with self-reports of intentions, then he should be happy with self-reports of motives.

Greenwald might complain that an ambiguity remains: on my understanding of the conative component, measures of behavior (e.g., observations of overt action) don't test (b) relations of the conative to other attitude components, but do they test (a) the attitude-behavior relationship or (c) the relation between behavior and the conative component? I reply that it depends on what *else* is measured besides behavior: if it's only motives, then the answer is (c), but if it's also affective and cognitive responses, then the answer is (a).

My response to Greenwald is not the only possible way of clarifying the attitude-behavior relationship: Ajzen (1989) proposes another way. Ajzen suggests that the relationship between attitudes and behavior "is a question of what we say versus what we

do" (1989: 244), so that "most tests of the attitude-behavior relation are better conceptualized as tests of the relation between verbal and nonverbal indicators of the same evaluative attitude" (1989: 245); it just so happens that typically one uses "evaluative responses of a cognitive or affective nature on the verbal side and evaluative responses of a conative kind on the nonverbal side" (1989: 244-5). I find Ajzen's suggestion unsatisfactory because it fails to take into account the fact that attitudes may be *faked*, so that what we say may not be what is the case. It may be true that *in practice* the best available tests of the attitude-behavior relationship compare nonverbal behavioral responses with verbal affective and cognitive responses; but it doesn't follow that the attitude-behavior relationship should be *conceptualized* as a question of what we say versus what we do rather than as a question of what we feel (believe, desire) versus what we do.[2]

2. Objections from unclarity and unparsimoniousness

How exactly is an attitude related to its components? Traditional tripartite definitions of attitudes (e.g., Rosenberg & Hovland 1960) may be thought to be unclear on this point. Two possibilities come to mind. A first possibility is that the three components are "three aspects of a single attitude" (McGuire 1989: 41), in the sense that an attitude is an "evaluative disposition" which is "manifested" by affective, cognitive, and conative responses. The idea here is that "correlations between responses of different classes are positive because these responses are manifestations of a position on a common underlying evaluative continuum" (Eagly & Chaiken 1993: 14). A second

---

[2] My understanding of the conative component as a disposition to have certain motives enables one to dispose easily of another objection to tripartite definitions raised by Zanna and Rempel, namely the objection that such definitions resolve the empirical question of attitude-behavior consistency by "definitional necessity": "the three-component view ... tends to prejudge the attitude-behavior relation, assuming that, almost by definition, such a relation must exist" (1988: 316). No. The motives that correspond to some attitudes may be so weak that they never result in behavior: recall the example in which I dislike my job but I never quit.

possibility is that "the three components are defined independently and yet comprise, at a higher level of abstraction, the single construct of attitude" (Ajzen 1989: 245). If the first possibility is the case, then it seems that an attitude should be defined as an evaluative disposition, and one should say only as a corollary that an attitude consists of three components. If the second possibility is the case, then why speak of attitudes at all, rather than just speaking of affective, cognitive, and conative dispositions? Speaking of attitudes would be like, e.g., speaking of "chair-desk-lamps"—motley amalgamations (mereological sums) of distinct and only loosely related entities. In either case, a tripartite definition of attitudes is problematic: either it states something that is a corollary of what should be the definition, or it is unparsimonious because it introduces a redundant concept of attitude. Or so the objection goes.

Concerning the first possibility, it is probably not the case. Even if *usually* a common evaluative continuum underlies affective, cognitive, and conative responses, this need not *always* be so: some attitudes may be *fragmented*. For example, I may think very highly of you because of your integrity but be unable to control my negative affective responses to you because you are a homosexual. Whether fragmented attitudes exist is an empirical question, but given that the possibility of their existence is intuitively obvious, it seems unwise to exclude this possibility by definitional fiat—i.e., by defining attitudes as evaluative dispositions. One might object that defining attitudes as evaluative dispositions does not exclude the possibility that fragmented attitudes exist: "High intercomponent correlations do not necessarily follow from the tripartite view. For example, affect, behavior, and cognition can sometimes be the product of very different learning situations" (Breckler 1984: 1193). I reply that I call an attitude 'fragmented' not when the intercomponent correlations are not high, but rather when there is no common underlying evaluative continuum. One might respond that my distinction is empirically vacuous: when intercomponent correlations are low, how could one distinguish empirically cases in which an underlying evaluative continuum exists from cases in

which it doesn't? I reply that the possibility that fragmented attitudes exist has the empirical consequence that in some cases intercomponent correlations will be *negative* (not just low), a consequence that seems to be excluded by the definition of attitudes as evaluative dispositions.

Concerning the second possibility, I claim that the concept of attitude need not be redundant even when no common evaluative continuum underlies affective, cognitive, and conative responses. The reason is that it's frequently useful to group together our (dispositional) responses towards certain objects. It's useful, for instance, to have a concept summarizing one's responses towards snakes, so there is reason for speaking of attitudes towards snakes, rather than just speaking of affective, cognitive, and conative dispositions towards snakes, even if the three dispositions reflect no common underlying evaluative continuum. One might object by remarking that we don't always group together *all* of our responses towards an object: one can have more than one distinct attitudes towards the *same* object (e.g., one may both admire and dislike a brilliant but obnoxious person). I reply that this remark indicates that about *any* set of responses towards an object could be considered an attitude; thus this remark actually supports understanding my tripartite definition as *not requiring* (though not excluding either) that a common evaluative continuum underlie the three components.[3]

3. Objections from empirical studies

Tripartite definitions of attitudes presuppose that a distinction can be drawn between affective, cognitive, and conative responses. Eagly and Chaiken (1993) grant

---

[3] A different charge of unparsimoniousness is raised by Cacioppo et al., who claim that a "possibly undesirable feature of the tripartite model ... is the unparsimonious notion that affective, cognitive, and behavioral stimuli can influence affective, cognitive, and behavioral attitude-components, each of which in turn can mediate affective, cognitive, and behavioral responses" (1989: 292). This charge does not apply to my tripartite definition, however, because my definition is neutral on questions like whether affective stimuli can influence cognitive responses.

that "the distinction must be accorded a certain heuristic value" (cf. Hilgard 1980) but claim that, "to be worth preserving in modern attitude theory, the distinction should have more than heuristic value": it "must have some discriminant validity". That is, "responses within each of the three categories should relate more strongly to other responses within that category than to responses in the other two categories" (1993: 12). Issues related to discriminant (and convergent) validity have been examined in several empirical studies (Bagozzi & Burnkrant 1979, 1985; Breckler 1984; Dillon & Kumar 1985; Fishbein & Ajzen 1974; Khotandapani 1971; Mann 1959; Mendler et al. 1990; Ostrom 1969; Van de Ven, Bornholt, & Bailey 1996; Widaman 1985; Woodmansee & Cook 1967), which have, however, yielded mixed results. After reviewing some of the relevant studies, Eagly and Chaiken conclude: "it appears that a definitive empirical determination of the dimensionality of evaluative responses is unlikely in the near future" (1993: 12). Then they claim (1993: 13-4):

> Because cognitive, affective, and behavioral responses are often not empirically distinguishable as three classes, the three-component terminology is overly strong and is inappropriate in its implication that the three types of responses are generally distinct, that is, distinguishable in most people most of the time ... A formal three-component model will probably be rejected for many, perhaps even most, attitudes.

Note that Eagly and Chaiken don't claim that "many, perhaps even most, attitudes" have one or two (rather than three) components; this possibility is consistent with my tripartite definition, which claims that an attitude has three components *in general*, not *always*. Eagly and Chaiken claim rather that in many cases one is unable to distinguish empirically between three classes of responses. This inability, however, need not reflect a problem with tripartite definitions: it may be just due to measurement problems. Such problems certainly exist: although a test of the discriminant validity of the affective/cognitive/conative distinction "requires an a priori method for classifying measures of affect, behavior, and cognition" (Breckler 1984: 1194), there is

disagreement about the appropriate classification of some measures. For example, the semantic differential is considered by Bagozzi and Burnkrant (1979: 918-9) "to tap the affective dimensions of attitudes", but is considered by Breckler (1984: 1198) "a measure of cognition" (cf. Ajzen 1989: 246). As another example, I argued in §1 that it's inappropriate to consider self-reports of intentions and endorsements of statements about hypothetical actions as measures of the conative component. I will give now two reasons for believing that the (putative) inability to distinguish empirically between affective, cognitive, and conative responses is an artifact of measurement.

First: most empirical studies have used exclusively verbal measures of affective, cognitive, and conative responses. However, as Bagozzi & Burnkrant (1985: 49) point out,

> it is in general a difficult task to demonstrate discriminant validity between measures of similar constructs using the same method of measurement. That is, because all measures were self-report indicators, it would be expected that any shared methods variation across measures would press for an artificial convergence and thus reduce discrimination. The evidence for discriminant validity in the face of such effects is thus strong evidence. It is a much easier task to demonstrate discriminant validity when maximally dissimilar methods are used.

In fact, Breckler (1984), who is perhaps the only investigator who used nonverbal measures, easily established discriminant validity.

The second reason is more speculative but also more interesting. Cacioppo et al. (1989: 293) point out that, in most relevant studies,

> the indices of "cognition," "affect," and "behavior" have been scaled to reflect *evaluations* of the attitude objects. For example, Breckler (1984) obtained thought listings about an attitude object; however, rather than using the total number of issue-relevant thoughts (or some other cognitive structure index) as a measure of the cognitive component, he had the subjects rate each thought along an evaluative dimension and used the ratio of the favorable to unfavorable thoughts about the attitude object as an index of the cognitive component.

Cacioppo et al. go on to conjecture that scaling the measures of independent dimensions along a common evaluative continuum may eliminate the empirical independence between the dimensions, though not to such a degree that the dimensions become isomorphic (1989: 294). In support of this conjecture they refer to a study in which measures of the "theoretically and empirically orthogonal" dimensions of activity and potency, after being transformed by such a scaling process, reflected "in part the original dimension of meaning and in part a new and common dimension" (1989: 294). If this conjecture is true, then it may be the case that the difficulty in establishing discriminant validity in tests of tripartite definitions of attitudes is partly due to the above scaling process.[4] This possibility suggests a direction for future research: one could reanalyze the data of previous studies and conduct new studies of discriminant validity by using *unscaled* measures.

I have defended my tripartite definition of attitudes against a number of attacks. My defense is partly programmatic, but for the moment one may conclude that a harsh evaluation of my tripartite definition is *not* warranted.

---

[4] Cacioppo et al. (1989) use their conjecture to *criticize* tripartite definitions of attitudes. Their reasoning is unclear to me, but maybe their point is that *convergent* (rather than lack of discriminant) validity is an artifact of the scaling process. This possibility creates no problem for my tripartite definition: as I explained in §2, my definition does not presuppose that the correlations between the three components are high—or even positive.

REFERENCES

Aderman, D., & Berkowitz, L. (1970). Observational set, empathy, and helping. *Journal of Personality and Social Psychology, 14,* 141-148.

Ajzen, I. (1989). Attitude structure and behavior. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 241-274). Hillsdale, NJ: Erlbaum.

Ancona, L., & Pareyson, R. (1968). Contributo allo studio della aggressione: La dinamica della obbedienza distruttiva [Contribution to the study of aggression: The dynamics of destructive obedience]. *Archivio di Psicologia, Neurologia, e Psichiatria, 29,* 340-372.

Anderson, J. (1974, May). *Bystander intervention in an assault.* Paper presented at the meeting of the Southeastern Psychological Association, Hollywood, FL.

Aristotle. (1985). *Nicomachean Ethics* (T. H. Irwin, Trans.). Indianapolis: Hackett.

Aronson, E., Ellsworth, P. C., Carlsmith, J. M., & Gonzales, M. H. (1990). *Methods of research in social psychology* (2nd ed.). New York: McGraw-Hill.

Athanassoulis, N. (2000). A response to Harman: Virtue ethics and character traits. *Proceedings of the Aristotelian Society, 100,* 215-221.

Atwell, J. E. (1982). Kant's notion of respect for persons. In O. H. Green (Ed.), *Respect for persons* (pp. 17-30). *Tulane Studies in Philosophy, 31.*

Austin, J. L. (1979). A plea for excuses. In J. O. Urmson & G. J. Warnock (Eds.), *J. L. Austin: Philosophical papers* (3rd ed., pp. 175-204). Oxford: Oxford University Press. (Original work published 1956)

Austin, W. (1979). Sex differences in bystander intervention in a theft. *Journal of Personality and Social Psychology, 37,* 2110-2120.

Axelrod, R. (1984). *The evolution of cooperation.* New York: Basic Books.

Bach, K., & Harnish, R. M. (1987). Relevant questions. *Behavioral and Brain Sciences, 10,* 711-712.

Badhwar, N. K. (1996). The limited unity of virtue. *Noûs, 30,* 306-329.

Bagozzi, R. P., & Burnkrant, R. E. (1979). Attitude organization and the attitude-behavior relationship. *Journal of Personality and Social Psychology, 37,* 913-929.

Bagozzi, R. P., & Burnkrant, R. E. (1985). Attitude organization and the attitude-behavior relation: A reply to Dillon and Kumar. *Journal of Personality and Social Psychology, 49,* 47-57.

Banuazizi, A., & Movahedi, S. (1975). Interpersonal dynamics in a simulated prison: A methodological analysis. *American Psychologist, 30,* 152-160.

Batson, C. D. (1976). Latent aspects of "From Jerusalem to Jericho". In M. P. Golden (Ed.), *The research experience* (pp. 205-214). Itasca, IL: Peacock.

Batson, C. D. (1991). *The altruism question: Toward a social-psychological answer.* Hillsdale, NJ: Erlbaum.

Batson, C. D., Cochran, P. J., Biederman, M. F., Blosser, J. L., Ryan, M. J., & Vogt, B. (1978). Failure to help when in a hurry: Callousness or conflict? *Personality and Social Psychology Bulletin, 4,* 97-101.

Baumeister, R. F. (1997). *Evil: Inside human violence and cruelty.* New York: Freeman.

Baumeister, R. F. (2001, April). Violent pride: Do people turn violent because of self-hate, or self-love? *Scientific American, 284,* 96-101.

Baumeister, R. F., Smart, L. & Boden, J. M. (1996). Relation of threatened egotism to violence and aggression: The dark side of high self-esteem. *Psychological Review, 103,* 5-33.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin, 115,* 243-267.

Baumeister, R. F., Tice, D. M., & Hutton, D. G. (1989). Self-presentational motivations and personality differences in self-esteem. *Journal of Personality, 57,* 547-579.

Baumeister, R. F., & Wotman, S. R. (1992). *Breaking hearts: The two sides of unrequited love.* New York: Guilford.

Benson, P. (1987). Moral worth. *Philosophical Studies, 51,* 365-382.

Berkowitz, L., & Connor, W. H. (1966). Success, failure, and social responsibility. *Journal of Personality and Social Psychology, 4,* 664-669.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist, 37,* 245-257.

Bickman, L. (1979). Interpersonal influence and the reporting of a crime. *Personality and Social Psychology Bulletin, 5,* 32-35.

Bickman, L., & Green, S. K. (1977). Situational cues and crime reporting: Do signs make a difference? *Journal of Applied Social Psychology, 7,* 1-18.

Bickman, L., & Rosenbaum, D. P. (1977). Crime reporting as a function of bystander encouragement, surveillance, and credibility. *Journal of Personality and Social Psychology, 35,* 577-586.

Bierbrauer, G. (1979). Why did he do it? Attribution of obedience and the phenomenon of dispositional bias. *European Journal of Social Psychology, 9,* 67-84.

Billings, A. (1979). Conflict resolution in distressed and nondistressed married couples. *Journal of Consulting and Clinical Psychology, 47,* 368-376.

Birnbaum, M. H. (1972). Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology, 93,* 35-42.

Birnbaum, M. H. (1973). Morality judgment: Test of an averaging model with differential weights. *Journal of Experimental Psychology, 99,* 395-399.

Blass, T. (1991). Understanding behavior in the Milgram obedience experiment: The role of personality, situations, and their interactions. *Journal of Personality and Social Psychology, 60,* 398-413.

Blass, T. (1992). The social psychology of Stanley Milgram. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 25, pp. 277-329). New York: Academic Press.

Blass, T. (1996a). The Milgram obedience experiment: Support for a cognitive view of defensive attribution. *Journal of Social Psychology, 136,* 407-410.

Blass, T. (1996b). Attribution of responsibility and trust in the Milgram obedience experiment. *Journal of Applied Social Psychology, 26,* 1529-1535.

Blass, T. (1999). The Milgram paradigm after 35 years: Some things we now know about obedience to authority. *Journal of Applied Social Psychology, 29,* 955-978.

Blass, T. (Ed.). (2000). *Obedience to authority: Current perspectives on the Milgram paradigm.* Mahwah, NJ: Erlbaum.

Blum, L. A. (1994). *Moral perception and particularity.* New York: Cambridge University Press.

Blumberg, S. R., & Hokanson, J. E. (1983). The effects of another person's response style on interpersonal behavior in depression. *Journal of Abnormal Psychology, 92,* 196-209.

Bok, D. C., & Warren, N. C. (1972). Religious belief as a factor in obedience to destructive commands. *Review of Religious Research, 13,* 185-191.

Bok, H. (1996). Acting without choosing. *Noûs, 30,* 174-196.

Bond, E. J. (1983). *Reason and Value*. New York: Cambridge University Press.

Boxill, B. R. (1995). Self-respect and protest. In R. S. Dillon (Ed.), *Dignity, character, and self-respect* (pp. 93-104). New York: Routledge. (Original work published 1976)

Brandt, R. B. (1992a). A utilitarian theory of excuses. In R. B. Brandt, *Morality, utilitarianism, and rights* (pp. 215-234). New York: Cambridge University Press. (Original work published 1969)

Brandt, R. B. (1992b). Traits of character: A conceptual analysis. In R. B. Brandt, *Morality, utilitarianism, and rights* (pp. 263-288). New York: Cambridge University Press. (Original work published 1970)

Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology, 47,* 1191-1205.

Brink, D. O. (1997). Kantian rationalism: Inescapability, authority, and supremacy. In G. Cullity & B. Gaut (Eds), *Ethics and practical reason* (pp. 255-291). Oxford: Clarendon.

Broadie, A. & Pybus, E. M. (1975). Kant's concept of "respect". *Kant-Studien, 66,* 58-64.

Brody, B. A. (1982). Towards a theory of respect for persons. In O. H. Green (Ed.), *Respect for persons* (pp. 61-76). *Tulane Studies in Philosophy, 31.*

Browning, C. R. (1992). *Ordinary men: Reserve police battalion 101 and the final solution in Poland.* New York: HarperCollins.

Burley, P. M., & McGuinness, J. (1977). Effects of social intelligence on the Milgram paradigm. *Psychological Reports, 40,* 767-770.

Bushman, B. J., & Baumeister, R. F. (1998). Threatened egotism, narcissism, self-esteem and direct and displaced aggression: Does self-love or self-hate lead to violence? *Journal of Personality and Social Psychology, 75,* 219-229.

Buss, S. (1999a). Appearing respectful: The moral significance of manners. *Ethics, 109,* 795-826.

Buss, S. (1999b). Respect for persons. *Canadian Journal of Philosophy, 29,* 517-550.

Cacioppo, J. T., Petty, R. E., & Geen, T. R. (1989). Attitude structure and function: From the tripartite to the homeostatic model of attitudes. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 275-309). Hillsdale, NJ: Erlbaum.

Campbell, J. (1999). Can philosophical accounts of altruism accommodate experimental data on helping behaviour? *Australasian Journal of Philosophy, 77,* 26-45.

Carlson, M., Charlin, V., & Miller, N. (1988). Positive mood and helping behavior: A test of six hypotheses. *Journal of Personality and Social Psychology, 55,* 211-229.

Carlson, M., & Miller, N. (1987). Explanation of the relation between negative mood and helping. *Psychological Bulletin, 102,* 91-108.

Chang, R. (Ed.). (1997). *Incommensurability, incomparability, and practical reason.* Cambridge, MA: Harvard University Press.

Chisholm, R. M. (1963). Supererogation and offense. *Ratio, 5,* 3-15.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin, 125,* 47-63.

Clark, R. D., III, & Word, L. E. (1972). Why don't bystanders help? Because of ambiguity? *Journal of Personality and Social Psychology, 24,* 392-400.

Clark, R. D., III, & Word, L. E. (1974). Where is the apathetic bystander? Situational characteristics of the emergency. *Journal of Personality and Social Psychology, 29,* 279-287.

Constanzo, E. M. (1976). *The effect of probable retaliation and sex related variables on obedience.* Doctoral dissertation, University of Wyoming.

Coutts, L. M. (1977). A note on Mixon's critique of Milgram's obedience research. *Personality and Social Psychology Bulletin, 3,* 519-521.

Cranor, C. (1975). Toward a theory of respect for persons. *American Philosophical Quarterly, 12,* 309-319.

Cranor, C. (1980). Kant's respect-for-persons principle. *International Studies in Philosophy, 12,* 19-39.

Cranor, C. (1982). Limitations of respect-for-persons theories. In O. H. Green (Ed.), *Respect for persons* (pp. 45-60). *Tulane Studies in Philosophy, 31.*

Cranor, C. (1983). On respecting human beings as persons. *Journal of Value Inquiry, 17,* 103-117.

Crowder, J. E. (1972). Relationship between therapist and client interpersonal behaviors and psychotherapy outcome. *Journal of Counseling Psychology, 19,* 68-75.

Cullity, G. (1995). Moral character and the iteration problem. *Utilitas, 7,* 289-299.

Cusumano, D. R., & Richey, M. H. (1970). Negative salience in impressions of character: Effects of extremeness of stimulus information. *Psychonomic Science, 20,* 81-83.

Daly, M., & Wilson, M. (1988). *Homicide.* New York: De Gruyter.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York: Avon.

Darley, J. M. (1995). Constructive and destructive obedience: A taxonomy of principal-agent relationships. *Journal of Social Issues, 51,* 125-154.

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27,* 100-108.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8,* 377-383.

Darley, J. M., Teger, A. I., & Lewis, L. D. (1973). Do groups always inhibit individuals' responses to emergencies? *Journal of Personality and Social Psychology, 26,* 395-399.

Darwall, S. L. (1995). Two kinds of respect. In R. S. Dillon (Ed.), *Dignity, character, and self-respect* (pp. 181-197). New York: Routledge. (Original work published 1977)

Deigh, J. (1995). Shame and self-esteem: A critique. In R. S. Dillon (Ed.), *Dignity, character, and self-respect* (pp. 133-156). New York: Routledge. (Original work published 1983)

DeJong, W. (1975). Another look at Banuazizi and Movahedi's analysis of the Stanford Prison Experiment. *American Psychologist, 30,* 1013-1015.

DeJong, W., Marber, S., & Shaver, R. (1980). Crime intervention: The role of a victim's behavior in reducing situational ambiguity. *Personality and Social Psychology Bulletin, 6,* 113-118.

Denner, B. (1968). Did a crime occur? Should I inform anyone? A study of deception. *Journal of Personality, 36,* 454-465.

DePaul, M. (2000). Character traits, virtues, and vices: Are there none? In B. Elevitch (Ed.), *Proceedings of the 20th World Congress of Philosophy: Vol. 9. Philosophy of mind* (pp. 141-157). Bowling Green, OH: Philosophy Documentation Center.

Dillon, R. S. (1992a). Respect and care: Toward moral integration. *Canadian Journal of Philosophy, 22,* 105-131.

Dillon, R. S. (1992b). How to lose your self-respect. *American Philosophical Quarterly, 29,* 125-139.

Dillon, R. S. (1992c). Toward a feminist conception of self-respect. *Hypatia, 7,* 52-69.

Dillon, R. S. (1992d). Care and respect. In E. Browning Cole & S. Coultrap-McQuin (Eds.), *Explorations in feminist ethics: Theory and practice* (pp. 69-81). Bloomington: Indiana University Press.

Dillon, R. S. (1995). (Ed.). *Dignity, character, and self-respect.* New York: Routledge.

Dillon, R. S. (1997). Self-respect: Moral, emotional, political. *Ethics, 107,* 226-249.

Dillon, W. R., & Kumar, A. (1985). Attitude organization and the attitude-behavior relation: A critique of Bagozzi and Burnkrant's reanalysis of Fishbein and Ajzen. *Journal of Personality and Social Psychology, 49,* 33-46.

Donagan, A. (1977). *The theory of morality.* Chicago: University of Chicago Press.

Doris, J. M. (1996). *People like us: Morality, psychology, and the fragmentation of character.* Doctoral dissertation, University of Michigan.

Doris, J. M. (1998). Persons, situations, and virtue ethics. *Noûs, 32,* 504-530.

Doris, J. M. (in press). *Lack of character: Personality and moral behavior.* New York: Cambridge University Press.

Douglas, J., & Olshaker, M. (1995). *Mindhunter: Inside the FBI's elite serial crime unit.* New York: Pocket.

Downie, R. S., & Telfer, E. (1970). *Respect for persons.* New York: Schocken.

Doyle, C. L. (1975). Interpersonal dynamics in role playing. *American Psychologist, 30,* 1011-1013.

Dudycha, G. J. (1936). An objective study of punctuality in relation to personality and achievement. *Archives of Psychology, 29,* 1-53.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth: Harcourt, Brace, Jovanovich.

Edel, A. (1974). The place of respect for persons in moral philosophy. *Philosophy in Context, 3*, 23-32.

Edwards, D. M., Franks, P., Friedgood, D., Lobban, G., & Mackay, H. C. G. (1969). *An experiment on obedience.* Unpublished student report, University of the Witwatersrand, Johannesburg, South Africa.

Epstein, S. (1977). Traits are alive and well. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 83-98). Hillsdale, NJ: Erlbaum.

Epstein, S. (1979a). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37,* 1097-1126.

Epstein, S. (1979b). Explorations in personality today and tomorrow: A tribute to Henry A. Murray. *American Psychologist, 34,* 649-653.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35,* 790-806.

Epstein, S. (1983a). The stability of confusion: A reply to Mischel and Peake. *Psychological Review, 90,* 179-184.

Epstein, S. (1983b). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360-392.

Epstein, S. (1984). The stability of behavior across time and situations. In R. Zucker, J. Aronoff, & A. I. Rabin (Eds.), *Personality and the prediction of behavior* (pp. 209-268). San Diego, CA: Academic Press.

Epstein, S. (1986). Does aggregation produce spuriously high estimates of behavior stability? *Journal of Personality and Social Psychology, 50,* 1199-1210.

Evans, R. I. (Ed.). (1980). *The making of social psychology: Discussions with creative contributors*. New York: Gardner.

Faber, N. (1971, October 15). "I almost considered the prisoners as cattle". *Life Magazine, 71,* 82-3.

Falls, M. M. (1987). Retribution, reciprocity, and respect for persons. *Law and Philosophy, 6,* 25-51.

Farrington, D. P. (1979). Experiments on deviance with special reference to dishonesty. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 12, pp. 207-252). New York: Academic Press.

Feinberg, J. (1970a). Supererogation and rules. In J. Feinberg, *Doing and deserving: Essays in the theory of responsibility* (pp. 3-24). Princeton: Princeton University Press. (Original work published 1961)

Feinberg, J. (1970b). Justice and personal desert. In J. Feinberg, *Doing and deserving: Essays in the theory of responsibility* (pp. 55-87). Princeton: Princeton University Press. (Original work published 1963)

Feinberg, J. (1973). Some conjectures about the concept of respect. *Journal of Social Philosophy, 4,* 1-3.

Fishbein, M., & Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review, 81,* 59-74.

Fisher, R., & Brown, S. (1988). *Getting together: Building relationships as we negotiate.* New York: Penguin.

Flanagan, O. (1991). *Varieties of moral personality: Ethics and psychological realism.* Cambridge, MA: Harvard University Press.

Flynn, C. P. (1977). *Insult and society: Patterns of comparative interaction.* Port Washington, NY: Kennikat.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge, MA: MIT Press.

Frank, J. D. (1944). Experimental studies of personal pressure and resistance: I. Experimental production of resistance. *Journal of General Psychology, 30,* 23-41.

Frankena, W. K. (1986). The ethics of respect for persons. *Philosophical Topics, 14,* 149-167.

Frankfurt, H. G. (1986). Freedom of the will and the concept of a person. In J. M. Fischer (Ed.), *Moral responsibility* (pp. 65-80). Ithaca: Cornell University Press. (Original work published 1971)

Funder, D. C., & Ozer, D. J. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology, 44,* 107-112.

Gaertner, S. L. (1975). The role of racial attitudes in helping behavior. *Journal of Social Psychology, 97,* 95-101.

Gaertner, S. L., & Dovidio, J. F. (1977). The subtlety of white racism, arousal, and helping behavior. *Journal of Personality and Social Psychology, 35,* 691-707.

Gauthier, D. (1986). *Morals by agreement.* New York: Oxford University Press.

Gelfand, D. M., Hartman, D. P., Walder, P., & Page, B. (1973). Who reports shoplifters? An experimental study. *Journal of Personality and Social Psychology, 25,* 276-285.

Ghosh-Dastidar, K. (1987). Respect for persons and self-respect: Western and Indian. *Journal of Indian Council of Philosophical Research, 5,* 83-93.

Gibbard, A. F. (1990a). *Utilitarianism and coordination.* Doctoral dissertation, Harvard University. New York: Garland. (Original work published 1971)

Gibbard, A. F. (1990b). *Wise choices, apt feelings: A theory of normative judgment.* Cambridge, MA: Harvard University Press.

Gidron, D., Koehler, D. J., & Tversky, A. (1993). Implicit quantification of personality traits. *Personality and Social Psychology Bulletin, 19,* 594-604.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117,* 21-38.

Gilbert, S. J. (1981). Another look at the Milgram obedience studies: The role of the gradated series of shocks. *Personality and Social Psychology Bulletin, 7,* 690-695.

Gourevitch, P. (1998). *We wish to inform you that tomorrow we will be killed with our families: Stories from Rwanda.* New York: Picador.

Greenberg, J. (1990). Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology, 75,* 561-568.

Greenwald, A. G. (1989a). Why are attitudes important? In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 1-10). Hillsdale, NJ: Erlbaum.

Greenwald, A. G. (1989b). Why attitudes are important: Defining attitude and attitude theory 20 years later. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 429-440). Hillsdale, NJ: Erlbaum.

Gregor, M. J. (1963). *Laws of freedom.* Oxford: Blackwell.

Grice, H. P. (1989a). Utterer's meaning and intentions. In H. P. Grice, *Studies in the way of words* (pp. 86-116). Cambridge: Harvard University Press. (Original work published 1969)

Grice, H. P. (1989b). Meaning. In H. P. Grice, *Studies in the way of words* (pp. 213-223). Cambridge: Harvard University Press. (Original work published 1957)

Grice, H. P. (1989c). Meaning revisited. In H. P. Grice, *Studies in the way of words* (pp. 283-303). Cambridge: Harvard University Press. (Original work published 1982)

Griffin, J. (1986). *Well-being: Its meaning, measurement and moral importance.* Oxford: Clarendon.

Gruzalski, B. (1982). Two accounts of our obligations to respect persons. In O. H. Green (Ed.), *Respect for persons* (pp. 77-89). *Tulane Studies in Philosophy, 31.*

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hamilton, V. L. (1992). Thoughts on obedience: A social structural view. *Contemporary Psychology, 37,* 1313.

Hampshire, S. (1953). Dispositions. *Analysis, 14,* 5-11.

Haney, C. (1976). The play's the thing: Methodological notes on social simulations. In M. P. Golden (Ed.), *The research experience* (pp. 177-190). Itasca, IL: Peacock.

Haney, C., Banks, C., & Zimbardo, P. G. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology, 1,* 69-97.

Haney, C., Banks, C., & Zimbardo, P. G. (1976). Interpersonal dynamics in a simulated prison. In M. P. Golden (Ed.), *The research experience* (pp. 157-177). Itasca, IL: Peacock.

Haney, C., & Zimbardo, P. G. (1977). The socialization into criminality: On becoming a prisoner and a guard. In J. L. Tapp & F. J. Levine (Eds.), *Law, justice, and the individual in society* (pp. 198-223). New York: Holt, Rinehart and Winston.

Haney, C., & Zimbardo, P. G. (1998). The past and future of U.S. prison policy: Twenty-five years after the Stanford Prison Experiment. *American Psychologist, 53,* 709-727.

Harari, H., Harari, O., & White, R. V. (1985). The reaction to rape by American male bystanders. *Journal of Social Psychology, 125,* 653-658.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society, 99,* 315-331.

Harman, G. (2000). The nonexistence of character traits. *Proceedings of the Aristotelian Society, 100,* 223-226.

Harré, R. (1979). *Social being: A theory for social psychology.* Oxford: Blackwell.

Harris, E. E. (1966). Respect for persons. In R. T. De George (Ed.), *Ethics and society: Original essays on contemporary moral problems* (pp. 111-132). Garden City, NY: Anchor.

Harris, V. A., & Robinson, C. E. (1973). Bystander intervention: Group size and victim status. *Bulletin of the Psychonomic Society, 2,* 8-10.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: I. Studies in deceit.* New York: Macmillan.

Hartshorne, H., May, M. A., & Maller, J. B. (1929). *Studies in the nature of character: II. Studies in service and self-control.* New York: Macmillan.

Hartshorne, H., May, M. A., & Shuttleworth, F. K. (1930). *Studies in the nature of character: III. Studies in the organization of character.* New York: Macmillan.

Henson, R. G. (1979). What Kant might have said: Moral worth and the overdetermination of dutiful action. *Philosophical Review, 88,* 39-54.

Herman, B. (1981). On the value of acting from the motive of duty. *Philosophical Review, 90,* 359-382.

Hicks, D. C. (1971). Respect for persons and respect for living things. *Philosophy, 46,* 346-348.

Hilgard, E. R. (1980). The trilogy of mind: Cognition, affection, and conation. *Journal of the History of the Behavioral Sciences, 16,* 107-117.

Hill, T. E., Jr. (1971). Kant on imperfect duty and supererogation. *Kant-Studien, 62,* 55-76.

Hill, T. E., Jr. (1995a). Servility and self-respect. In R. S. Dillon (Ed.), *Dignity, character, and self-respect* (pp. 76-92). New York: Routledge. (Original work published 1973)

Hill, T. E., Jr. (1995b). Self-respect reconsidered. In R. S. Dillon (Ed.), *Dignity, character, and self-respect* (pp. 117-124). New York: Routledge. (Original work published 1985)

Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken, *The psychology of attitudes* (pp. 23-87). Fort Worth, TX: Harcourt Brace.

Hofling, C. K., Brotzman, E., Darlymple, S., Graves, N., & Pierce, C. M. (1966). An experimental study in nurse-physician relationships. *Journal of Nervous and Mental Disease, 143,* 171-180.

Horowitz, I. A. (1971). The effect of group norms on bystander intervention. *Journal of Social Psychology, 83,* 265-273.

Howard, W., & Crano, W. D. (1974). Effects of sex, conversation, location, and size of observer group on bystander intervention in a high risk situation. *Sociometry, 37,* 491-507.

Hudson, S. D. (1980). The nature of respect. *Social Theory and Practice, 6,* 69-90.

Hunt, M. 1990. *The compassionate beast: What science is discovering about the humane side of humankind.* New York: Morrow.

Ingram, P. (1979). Deception, obedience, and authority. *Philosophy, 54,* 529-533.

Irwin, T. H. (1988). Disunity in the Aristotelian virtues. In J. Annas & R. H. Grimm (Eds.), *Oxford studies in ancient philosophy: Supplementary volume* (pp. 61-78). Oxford: Clarendon.

Irwin, T. H. (Trans.). (1985). *Aristotle: Nicomachean Ethics.* Indianapolis: Hackett.

Isen, A. M. (1970). Success, failure, attention, and reaction to others: The warm glow of success. *Journal of Personality and Social Psychology, 15,* 294-301.

Isen, A. M., Clark, M., & Schwartz, M. F. (1976). Duration of the effect of good mood on helping: "Footprints on the sands of time". *Journal of Personality and Social Psychology, 34,* 385-393.

Isen, A. M., Horn, N., & Rosenhan, D. L. (1973). Effects of success and failure on children's generosity. *Journal of Personality and Social Psychology, 27,* 239-247.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology, 21,* 384-388.

Isen, A. M., Shalker, T. E., Clark, M., & Karp, L. (1978). Affect, accessibility of material in memory, and behavior: A cognitive loop? *Journal of Personality and Social Psychology, 36,* 1-12.

Isen, A. M., & Simmonds, S. F. (1978). The effect of feeling good on a helping task that is incompatible with good mood. *Sociometry, 41,* 346-349.

Kamm, F. M. (1999). *Does distance matter morally to the duty to rescue?* Unpublished manuscript.

Kanouse, D. E., & Hanson, L. R., Jr. (1972). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 47-62). Morristown, NJ: General Learning Press.

Kant, I. (1964). *Groundwork of the metaphysic of morals* (H. J. Paton, Trans.). New York: Harper & Row. (Original work published 1785 as *Grundlegung zur Metaphysik der Sitten*; 2nd edition 1786. Page references to volume 4 of the Prussian Academy 1903 edition also given.)

Kant, I. (1993). *Critique of practical reason* (L. W. Beck, Trans.). Englewood Cliffs, NJ: Macmillan. (Original work published 1788 as *Kritik der praktischen Vernunft*. Page references to volume 5 of the Prussian Academy 1908 edition also given.)

Kant, I. (1996). *The metaphysics of morals* (M. Gregor, Trans.). Cambridge, MA: Cambridge University Press. (Original work published 1797 as *Die Metaphysik der Sitten*. Page references to volume 6 of the Prussian Academy 1907 edition also given.)

Kavka, G. S. (1984). The reconciliation project. In D. Copp & D. Zimmerman (Eds.), *Morality, reason and truth: New essays on the foundations of ethics* (pp. 297-319). Totowa, NJ: Rowman & Allanheld.

Khotandapani, V. (1971). Validation of feeling, belief, and intention to act as three components of attitude and their contribution to prediction of contraceptive behavior. *Journal of Personality and Social Psychology, 19,* 321-333.

Kilham, W., & Mann, L. (1974). Level of destructive obedience as a function of transmitter and executant roles in the Milgram obedience paradigm. *Journal of Personality and Social Psychology, 29,* 696-702.

Kleinig, J. (1971). The concept of desert. *American Philosophical Quarterly, 8,* 71-78.

Kolnai, A. (1995). Dignity. In R. S. Dillon (Ed.), *Dignity, character, and self-respect* (pp. 53-75). New York: Routledge. (Original work published 1976)

Kudirka, N. Z. (1965). *Defiance of authority under peer influence*. Doctoral dissertation, Yale University.

Kupperman, J. J. (1991). *Character*. New York: Oxford University Press.

Kupperman, J. J. (2001). The indispensability of character. *Philosophy, 76,* 239-250.

LaFollette, H. (2000). Gun control. *Ethics, 110,* 263-281.

Landis, C. (1924). Studies of emotional reactions: II. General behavior and facial expression. *Journal of Comparative Psychology, 4,* 447-509.

Latané, B., & Darley, J. M. (1969). Bystander "apathy". *American Scientist, 57,* 244-268.

Latané, B., & Darley, J. M. (1970a). Social determinants of bystander intervention in emergencies. In J. Macaulay & L. Berkowitz (Eds.), *Altruism and Helping Behavior* (pp. 13-27). New York: Academic Press.

Latané, B., & Darley, J. M. (1970b). *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.

Latané, B., & Darley, J. M. (1976). *Help in a crisis: Bystander response to an emergency.* Morristown, NJ: General Learning Press.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin, 89,* 308-324.

Latané, B., Nida, S. A., & Wilson, D. W. (1981). The effects of group size on helping behavior. In J. P. Rushton & R. M. Sorrentino (Eds.), *Altruism and helping behavior: Social, personality and developmental perspectives* (pp. 287-313). Hillsdale, NJ: Erlbaum.

Latané, B., & Rodin, J. (1969). A lady in distress: Inhibiting effects of friends and strangers on bystander intervention. *Journal of Experimental Social Psychology, 5,* 189-202.

Lauener, H. (1981). Der systematische Stellenwert des Gefühls der Achtung in Kants Ethik [The systematic importance of the feeling of respect in Kant's ethics]. *Dialectica, 35,* 243-264.

Levin, P. F., & Isen, A. M. (1975). Further studies on the effect of feeling good on helping. *Sociometry, 38,* 141-147.

Lewicka, M., Czapinski, J., & Peeters, G. (1992). Positive-negative asymmetry or 'When the heart needs a reason'. *European Journal of Social Psychology, 22,* 425-434.

Lewis, D. K. (1969). *Convention: A philosophical study.* Cambridge: Harvard University Press.

Lewis, D. K. (1983). Truth in fiction. In D. K. Lewis, *Philosophical papers: Volume I* (pp. 261-280). Oxford: Oxford University Press. (Original work published 1978)

Lifton, R. J. (1986). *The Nazi doctors: Medical killing and the psychology of genocide.* New York: Basic Books.

Lombardi, L. G. (1983). Inherent worth, respect, and rights. *Environmental Ethics, 5,* 257-270.

Lovibond, S. H., & Adams, W. G. (1979). Reply to Dr. Morgan's comments on S. H. Lovibond, Mithiran, & W. G. Adams: 'The effects of three experimental prison

environments on the behaviour of non-convict volunteer subjects'. *Australian Psychologist, 14,* 286-287.

Lovibond, S. H., Mithiran, & Adams, W. G. (1979). The effects of three experimental prison environments on the behaviour of non-convict volunteer subjects. *Australian Psychologist, 14,* 273-285.

MacKenzie, M. H. (1968). *The interpersonal behavior of normal and clinic family members.* Doctoral dissertation, Michigan State University.

Maclagan, W. G. (1960a). Respect for persons as a moral principle—I. *Philosophy, 35,* 193-217.

Maclagan, W. G. (1960b). Respect for persons as a moral principle—II. *Philosophy, 35,* 289-305.

Mandelbaum, M. (1955). *The phenomenology of moral experience.* Glencoe, IL: Free Press.

Mann, J. H. (1959). The relationship between cognitive, affective, and behavioral aspects of racial prejudice. *The Journal of Social Psychology, 49,* 223-228.

Mantell, D. M. (1971). The potential for violence in Germany. *Journal of Social Issues, 27,* 101-112.

Mantell, D. M., & Panzarella, R. (1976). Obedience and responsibility. *British Journal of Social and Clinical Psychology, 25,* 239-245.

Marsh, H. W. (1986). Global self-esteem: Its relation to specific facets of self-concept and their importance. *Journal of Personality and Social Psychology, 51,* 1224-1236.

Marsh, H. W. (1993). Relations between global and specific domains of self: The importance of individual importance, certainty, and ideals. *Journal of Personality and Social Psychology, 65,* 975-992.

Marsh, H. W. (1995). A Jamesian model of self-investment and self-esteem: Comment on Pelham (1995). *Journal of Personality and Social Psychology, 69,* 1151-1160.

Marsh, H. W., & Parducci, A. (1978). Natural anchoring at the neutral point of category rating scales. *Journal of Experimental Social Psychology, 14,* 193-204.

Martijn, C., Spears, R., Van der Pligt, J., & Jakobs, E. (1992). Negativity and positivity effects in person perception and inference: Ability versus morality. *European Journal of Social Psychology, 22,* 453-463.

Martin, J., Lobb, B., Chapman, G. C., & Spillane, R. (1976). Obedience under conditions demanding self-immolation. *Human Relations, 29,* 345-356.

Massey, S. J. (1983). Kant on self-respect. *Journal of the History of Philosophy, 21,* 57-73.

Massey, S. J. (1995). Is self-respect a moral or a psychological concept? In R. S. Dillon (Ed.), *Dignity, character, and self-respect* (pp. 198-217). New York: Routledge. (Original work published 1983)

Maxwell, R. J. & Silverman, P. (1978). The nature of deference. *Current Anthropology, 19,* 151.

McGuire, W. J. (1989). The structure of individual attitudes and attitude systems. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 37-69). Hillsdale, NJ: Erlbaum.

Meeus, W. H., & Raaijmakers, Q. A. W. (1995). Obedience in modern society: The Utrecht studies. *Journal of Social Issues, 51,* 155-175.

Melzack, R., & Wall, P. D. (1988). *The challenge of pain.* London: Penguin.

Mendler, W., Doll, J., & Orth, B. (1990). Zur Konstruktvalidität in der Einstellungsmessung: Singuläre oder multiple Komponenten-Modelle [On construct validity in attitude measurement: Single- or multiple-component models]. *Zeitschrift für Sozialpsychologie, 21,* 238-251.

Merritt, M. (1999). *Virtue ethics and the social psychology of character.* Doctoral dissertation, University of California, Berkeley.

Merritt, M. (2000). Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice, 3,* 365-383.

Meyer, M. J. (1989). Dignity, rights, and self-control. *Ethics, 99,* 520-534.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371-378.

Milgram, S. (1965a). *Obedience* [Film]. New York University Film Library.

Milgram, S. (1965b). Some conditions of obedience and disobedience to authority. *Human Relations, 18,* 57-76.

Milgram, S. (1965c). Liberating effects of group pressure. *Journal of Personality and Social Psychology, 1,* 127-134.

Milgram, S. (1972). Interpreting obedience: Error and evidence. A reply to Orne and Holland. In A. G. Miller (Ed.), *The social psychology of psychological research* (pp. 138-154). New York: Free Press.

Milgram, S. (1974). *Obedience to authority: An experimental view.* New York: Harper & Row.

Milgram, S. (1983). Reflections on Morelli's 'Dilemma of obedience'. *Metaphilosophy, 14,* 190-194.

Miller, A. G. (1986). *The obedience experiments: A case study of controversy in social science.* New York: Praeger.

Miller, A. G. (1995). Constructions of the obedience experiments: A focus upon domains of relevance. *Journal of Social Issues, 51,* 33-53.

Miller, A. G., Gillen, B., Schenker, C., & Radlove, S. (1974). The prediction and perception of obedience to authority. *Journal of Personality, 42,* 23-42.

Miranda, F. S. B., Caballero, R. B., Gomez, M. N. G., & Zamorano, M. A. M. (1981). Obediencia a la autoridad [Obedience to authority]. *Psiquis, 2,* 212-221.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89,* 730-755.

Mixon, D. (1972). Instead of deception. *Journal for the Theory of Social Behaviour, 2,* 145-177.

Mixon, D. (1989). *Obedience and civilization: Authorized crime and the normativity of evil.* London: Pluto Press.

Modigliani, A., & Rochat, F. (1995). The role of interaction sequences and the timing of resistance in shaping obedience and defiance to authority. *Journal of Social Issues, 51,* 107-123.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379-387.

Morelli, M. F. (1983). Milgram's dilemma of obedience. *Metaphilosophy, 14,* 183-189.

Morgan, A. H. (1979). Comments on S. H. Lovibond, Mithiran, & W. G. Adams: 'The effects of three experimental prison environments on the behaviour of non-convict volunteer subjects'. *Australian Psychologist, 14,* 285-286.

Moriarty, T. (1975). Crime, commitment, and the responsive bystander: Two field experiments. *Journal of Personality and Social Psychology, 31,* 370-376.

Movahedi, S., & Banuazizi, A. (1975). Reply to comments. *American Psychologist, 30,* 1016-1018.

Mueller, W. J. (1969). Patterns of behavior and their reciprocal impact in the family and in psychotherapy. *Journal of Counseling Psychology Monographs, 16* (No. 2, Part 2).

Mueller, W. J., & Dilling, C. A. (1968). Therapist-client interview behavior and personality characteristics of therapists. *Journal of Projective Techniques and Personality Assessment, 32,* 281-288.

Musen, K., & Zimbardo, P. G. (1992). *Quiet rage: The Stanford prison study* [Video]. Stanford University.

Nagel, T. (1970). *The possibility of altruism.* Princeton: Princeton University Press.

Nagel, T. (1979). Moral luck. In T. Nagel, *Mortal questions* (pp. 24-38). New York: Cambridge University Press. (Original work published 1976)

Newcomb, T. M. (1929). *Consistency of certain extrovert-introvert behavior patterns in 51 problem boys.* New York: Columbia University, Teachers College, Bureau of Publications.

Nisbett, R. E. (1980). The trait construct in lay and professional psychology. In L. Festinger (Ed.), *Retrospections of social psychology.* New York: Oxford University Press.

Noggle, R. (1999). Kantian respect and particular persons. *Canadian Journal of Philosophy, 29,* 449-478.

Norman, R. (1989). Respect for persons, autonomy and equality. *Revue Internationale de Philosophie, 43,* 323-341.

Nozick, R. (1974). *Anarchy, state, and utopia.* New York: Basic Books.

O'Leary, C. J., Willis, F. N., & Tomich, E. (1970). Conformity under deceptive and nondeceptive techniques. *Sociological Quarterly, 11,* 87-93.

Orford, J. (1986). The rules of interpersonal complementarity: Does hostility beget hostility and dominance, submission? *Psychological Review, 93,* 365-377.

Orlando, N. J. (1973). The mock ward: A study in simulation. In O. Milton & R. G. Wahler (Eds.), *Behavior disorders: Perspectives and trends* (pp. 162-170). Philadelphia: Lippincott.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776-783.

Orne, M. T., & Evans, F. J. (1965). Social control in the psychological experiment: Anti-social behavior and hypnosis. *Journal of Personality and Social Psychology, 1,* 189-200.

Orne, M. T., & Holland, C. H. (1968). On the ecological validity of laboratory deceptions. *International Journal of Psychiatry, 6,* 282-293.

Ostrom, T. M. (1969). The relationship between the affective, behavioral, and cognitive components of attitude. *Journal of Experimental Social Psychology, 5,* 12-30.

Parducci, A. (1968, December). The relativism of absolute judgments. *Scientific American, 219,* 84-90.

Parfit, D. (1984). *Reasons and persons.* Oxford: Clarendon.

Paton, H. J. (1948). *The categorical imperative: A study in Kant's moral philosophy.* Chicago: University of Chicago Press.

Patten, S. C. (1977a). The case that Milgram makes. *Philosophical Review, 86,* 350-364.

Patten, S. C. (1977b). Milgram's shocking experiments. *Philosophy, 52,* 425-440.

Peake, P. K. (1982). *Searching for consistency: The Carleton student behavior study.* Doctoral dissertation, Stanford University.

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1,* 33-60.

Pelham, B. W. (1995a). Self-investment and self-esteem: Evidence for a Jamesian model of self-worth. *Journal of Personality and Social Psychology, 69,* 1141-1150.

Pelham, B. W. (1995b). Further evidence for a Jamesian model of self-worth: Reply to Marsh (1995). *Journal of Personality and Social Psychology, 69,* 1161-1165.

Pelham, B. W., & Swann, W. B., Jr. (1989). From self-conceptions to self-worth: On the sources and structure of global self-esteem. *Journal of Personality and Social Psychology, 57,* 672-680.

Pigden, C. R., & Gillet, G. R. (1996). Milgram, method and morality. *Journal of Applied Philosophy, 13,* 233-250.

Piliavin, I. M., Piliavin, J. A., & Rodin, J. (1975). Costs, diffusion, and the stigmatized victim. *Journal of Personality and Social Psychology, 32,* 429-438.

Piliavin, I. M., Rodin, J., & Piliavin, J. A. (1969). Good Samaritanism: An underground phenomenon? *Journal of Personality and Social Psychology, 13,* 289-299.

Piliavin, J. A., & Piliavin, I. M. (1972). Effect of blood on reactions to a victim. *Journal of Personality and Social Psychology, 23,* 353-361.

Piliavin, J. A., Dovidio, J. F., Gaertner, S. L., & Clark, R. D., III. (1981). *Emergency intervention.* New York: Academic Press.

Piliavin, J. A., Dovidio, J. F., Gaertner, S. L., & Clark, R. D., III. (1982). Responsive bystanders: The process of intervention. In V. J. Derlega & J. Grzelak (Eds.), *Cooperation and helping behavior: Theories and research* (pp. 279-304). New York: Academic Press.

Pojman, L. P., & McLeod, O. (1999). (Eds.). *What do we deserve? A reader on justice and desert.* New York: Oxford University Press.

Powers, P. C., & Geen, R. G. (1972). Effects of the behavior and the perceived arousal of a model on instrumental aggression. *Journal of Personality and Social Psychology, 23,* 175-183.

Pratkanis, A. R. (1989). The cognitive representation of attitudes. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 71-98). Hillsdale, NJ: Erlbaum.

Quine, W. V. (1976). Quantifiers and propositional attitudes. In W. V. Quine, *The ways of paradox and other essays* (pp. 185-196). Cambridge, MA: Harvard University Press. (Original work published 1955)

Railton, P. A. (1995). Made in the shade: Moral compatibilism and the aims of moral theory. In J. Couture & K. Nielsen (Eds.), *On the relevance of metaethics: New essays on metaethics. Canadian Journal of Philosophy, Supplementary Volume 21* (pp. 79-106). Calgary, Alberta, Canada: University of Calgary Press.

Rank, S. G., & Jacobson, C. K. (1977). Hospital nurses' compliance with medication overdose orders: A failure to replicate. *Journal of Health and Social Behavior, 18,* 188-193.

Raush, H. L., Dittman, A. T., & Taylor, T. J. (1959). The interpersonal behavior of children in residential treatment. *Journal of Abnormal and Social Psychology, 58,* 9-27.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Reath, A. (1989). Kant's theory of moral sensibility: Respect for the moral law and the influence of inclination. *Kant-Studien, 80,* 284-302.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86,* 61-79.

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition, 4,* 1-17.

Reeder, G. D., Henderson, D. J., & Sullivan, J. J. (1982). From dispositions to behaviors: The flip side of attribution. *Journal of Research in Personality, 16,* 355-375.

Reeder, G. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology, 44,* 736-745.

Rice, P. L. K. (1969). *The modification of interpersonal roles.* Doctoral dissertation, West Virginia University.

Richey, M. H., Bono, F. S., Lewis, H. V., & Richey, H. W. (1982). Selectivity of negative bias in impression formation. *Journal of Social Psychology, 116,* 107-118.

Richey, M. H., & Dwyer, J. D. (1970). Negative salience in impressions of character: Sex differences. *Psychonomic Science, 20,* 77-79.

Richey, M. H., Koenigs, R. J., Richey, H. W., & Fortin, R. (1975). Negative salience in impressions of character: Effects of unequal proportions of positive and negative information. *Journal of Social Psychology, 97,* 233-241.

Richey, M. H., McClelland, L., & Shimkunas, A. M. (1967). Relative influence of positive and negative information in impression formation and persistence. *Journal of Personality and Social Psychology, 3,* 322-327.

Richey, M. H., Richey, H. W., & Thieman, G. (1972). Negative salience in impression of character: Effects of new information on established relationships. *Psychonomic Science, 28,* 65-67.

Ring, K., Wallston, K., & Corey, M. (1970). Mode of debriefing as a factor affecting subjective reaction to a Milgram-type obedience experiment: An ethical inquiry. *Representative Research in Social Psychology, 1,* 67-88.

Riskey, D. R., & Birnbaum, M. H. (1974). Compensatory effects in moral judgment: Two rights don't make up for a wrong. *Journal of Experimental Psychology, 103,* 171-173.

Rochat, F., Maggioni, O., & Modigliani, A. (2000). The dynamics of obeying and opposing authority: A mathematical model. In T. Blass (Ed.), *Obedience to authority: Current perspectives on the Milgram paradigm* (pp. 161-192). Mahwah, NJ: Erlbaum.

Rochat, F., & Modigliani, A. (2000). Captain Paul Grueninger: The chief of police who saved Jewish refugees by refusing to do his duty. In T. Blass (Ed.), *Obedience to authority: Current perspectives on the Milgram paradigm* (pp. 91-110). Mahwah, NJ: Erlbaum.

Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, affective, and behavioral components of attitudes. In M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. P. Abelson, & J. W. Brehm (Eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 1-14). New Haven: Yale University Press.

Rosenhan, D. (1969). Some origins of concern for others. In P. Mussen, J. Langer, & M. Covington (Eds.), *Trends and issues in developmental psychology* (pp. 134-153). New York: Holt, Rinehart and Winston.

Rosenthal, A. M. (1964). *Thirty-eight witnesses.* New York: McGraw-Hill.

Rosenthal, R. (1965). The volunteer subject. *Human Relations, 18,* 389-406.

Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject.* New York: Wiley.

Rosnow, R. L. (1993). The volunteer problem revisited. In P. D. Bolanck (Ed.), *Interpersonal expectations: Theory, research, and applications* (pp. 418-437). New York: Cambridge University Press.

Rosnow, R. L., & Rosenthal, R. (1976). The volunteer subject revisited. *Australian Journal of Psychology, 28,* 97-108.

Ross, A. S. (1971). Effect of increased responsibility on bystander intervention: The presence of children. *Journal of Personality and Social Psychology, 19,* 306-310.

Ross, A. S., & Braband, J. (1973). Effect of increased responsibility on bystander intervention, II: The cue value of a blind person. *Journal of Personality and Social Psychology, 25,* 254-258.

Ross, L. D. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 10, pp. 173-220). New York: Academic Press.

Ross, L. D. (1988). Situationist perspectives on the obedience experiments. *Contemporary Psychology, 33,* 101-104.

Ross, L. D., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology.* New York: McGraw-Hill.

Ross, L. D., & Thomas, E. (1986). *Notes on regression, aggregation, and the statistics of personal prediction.* Unpublished manuscript, Stanford University.

Ross, L. D., & Thomas, E. (1987). *The statistics of cross-situational consistency and behavioral predictability.* Unpublished manuscript, Stanford University.

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology, 50,* 131-142.

Rule, A. (1989). *The stranger beside me.* New York: Signet.

Rutkowski, G. K., Gruder, C. L., & Romer, D. (1983). Group cohesiveness, social norms, and bystander intervention. *Journal of Personality and Social Psychology, 44,* 545-552.

Sabini, J., & Silver, M. (1982). *Moralities of everyday life.* New York: Oxford University Press.

Sabini, J., & Silver, M. (1983). Dispositional vs. situational interpretations of Milgram's obedience experiments: 'The fundamental attribution error'. *Journal for the Theory of Social Behaviour, 13,* 147-154.

Sachs, D. (1981). How to distinguish self-respect from self-esteem. *Philosophy & Public Affairs, 10,* 346-360.

Sachs, D. (1982). Self-respect and respect for others: Are they independent? In O. H. Green (Ed.), *Respect for persons* (pp. 109-128). *Tulane Studies in Philosophy, 31.*

Safer, M. A. (1980). Attributing evil to the subject, not the situation: Student reaction to Milgram's film on obedience. *Personality and Social Psychology Bulletin, 6,* 205-209.

Salovey, P., Mayer, J. D., & Rosenhan, D. L. (1991). Mood and helping: Mood as a motivator of helping and helping as a regulator of mood. In M. S. Clark (Ed.), *Prosocial behavior* (*Review of Personality and Social Psychology,* vol. 12, pp. 215-237). London: Sage.

Schaller, M., & Cialdini, R. B. (1990). Happiness, sadness, and helping: A motivational integration. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition, vol. 2* (pp. 265-296). New York: Guilford.

Schedler, G. (1982). A theory of collective responsibility and some applications. *Heythrop Journal, 23,* 395-412.

Schiffer, S. R. (1972). *Meaning.* Oxford: Clarendon Press.

Schuman, H., & Scott, J. (1987). Problems in the use of survey questions to measure public opinion. *Science, 236,* 957-959.

Schurz, G. (1985). Experimentelle Überprüfung des Zusammenhangs zwischen Persönlichkeitsmerkmalen und der Bereitschaft zum destruktiven Gehorsam gegenüber Autoritäten [Experimental examination of the relationship between personality characteristics and the readiness to destructive obedience to authorities]. *Zeitschrift für experimentelle und angewandte Psychologie, 32,* 160-177.

Schwartz, S. H., & Clausen, G. T. (1970). Responsibility, norms, and helping in an emergency. *Journal of Personality and Social Psychology, 16,* 299-310.

Schwartz, S. H., & Gottlieb, A. (1976). Bystander reactions to a violent theft: Crime in Jerusalem. *Journal of Personality and Social Psychology, 34,* 1188-1199.

Schwartz, S. H., & Gottlieb, A. (1980). Bystander anonymity and reactions to emergencies. *Journal of Personality and Social Psychology, 39,* 418-430.

Schwarz, L., Jennings, K., Petrillo, J., & Kidd, R. F. (1980). Role of commitments in the decision to stop a theft. *Journal of Social Psychology, 110,* 183-192.

Schwarz, N., Groves, R. M., & Schuman, H. (1998). Survey methods. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of Social Psychology* (4th ed., Vol. 1, pp. 143-179). New York: Random House.

Sedikides, C., & Skowronski, J. J. (1993). The self in impression formation: Trait centrality and social perception. *Journal of Experimental Social Psychology, 29,* 347-357.

Shaffer, D. R., Rogel, M., & Hendrick, C. (1975). Intervention in the library: The effect of increased responsibility on bystander's willingness to prevent a theft. *Journal of Applied Social Psychology, 5,* 303-319.

Shalala, S. R. (1974). *A study of various communication settings which produce obedience by subordinates to unlawful superior orders.* Doctoral dissertation, University of Kansas.

Shanab, M. E., & O'Neill, P. (1979). The effects of contrast upon compliance with socially undesirable requests in the door-in-the-face paradigm. *Canadian Journal of Behavioural Science, 11,* 236-244.

Shanab, M. E., & Yahya, K. A. (1977). A behavioral study of obedience in children. *Journal of Personality and Social Psychology, 35,* 530-536.

Shanab, M. E., & Yahya, K. A. (1978). A cross-cultural study of obedience. *Bulletin of the Psychonomic Society, 11,* 267-269.

Shannon, J., & Guerney, B., Jr. (1973). Interpersonal effects of interpersonal behavior. *Journal of Personality and Social Psychology, 26,* 142-150.

Shelton, G. A. (1982). *The generalization of understanding to behaviour: The role of perspective in enlightenment.* Doctoral dissertation, University of British Columbia.

Sher, G. (1987). *Desert.* Princeton: Princeton University Press.

Sheridan, C. L., & King, R. G., Jr. (1972). Obedience to authority with an authentic victim. *Proceeding of the 80th Annual Convention of the American Psychological Association,* 165-166.

Shotland, R. L., & Heinold, W. D. (1985). Bystander response to arterial bleeding: Helping skills, the decision-making process, and differentiating the helping response. *Journal of Personality and Social Psychology, 49,* 347-356.

Shotland, R. L., & Stebbins, C. A. (1980). Bystander response to rape: Can a victim attract help? *Journal of Applied Social Psychology, 10,* 510-527.

Shotland, R. L., & Straw, M. K. (1976). Bystander response to an assault: When a man attacks a woman. *Journal of Personality and Social Psychology, 34,* 990-999.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52,* 689-699.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105,* 131-142.

Skowronski, J. J., & Carlston, D. E. (1992). Caught in the act: When impressions based on highly diagnostic behaviours are resistant to contradiction. *European Journal of Social Psychology, 22,* 435-452.

Smith, H. (1991). Varieties of moral worth and moral credit. *Ethics, 101,* 279-303.

Smith, M. (1994). *The moral problem.* Oxford: Blackwell.

Smith, P. B., & Bond, M. H. (1993). *Social psychology across cultures: Analysis and perspectives.* New York: Simon & Schuster.

Smith, R. E., Smythe, L., & Lien, D. (1972). Inhibition of helping behavior by a similar or dissimilar nonreactive fellow bystander. *Journal of Personality and Social Psychology, 23,* 414-419.

Smith, R. E., Vanderbilt, K., & Callen, M. B. (1973). Social comparison and bystander intervention in emergencies. *Journal of Applied Social Psychology, 3,* 186-196.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford: Blackwell.

Sreenivasan, G. (2000). *Errors about errors: Virtue theory and trait attribution.* Unpublished manuscript.

Staub, E. (1970). A child in distress: The influence of age and number of witnesses on children's attempts to help. *Journal of Personality and Social Psychology, 14,* 130-140.

Staub, E. (1971). Helping a person in distress: The influence of implicit and explicit "rules" of conduct on children and adults. *Journal of Personality and Social Psychology, 17,* 137-144.

Staub, E. (1974). Helping a distressed person: Social, personality, and stimulus determinants. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 7, pp. 293-341). New York: Academic Press.

Stocker, M. (1979). Desiring the bad: An essay in moral psychology. *The Journal of Philosophy, 76,* 738-753.

Strawson, P. F. (1971). Intention and convention in speech acts. In J. R. Searle (Ed.), *The philosophy of language.* (pp. 23-38). Oxford: Oxford University Press. (Original work published 1964)

Swenson, C. H. (1967). Psychotherapy as a special case of dyadic interaction: Some suggestions for theory and research. *Psychotherapy: Theory, Research and Practice, 4,* 7-13.

Sytsma, S. E. (1993). The role of *Achtung* in Kant's moral theory. *Auslegung, 19,* 117-122.

Tafarodi, R. W., & Swann, W. B., Jr. (1995). Self-liking and self-competence as dimensions of global self-esteem: Initial validation of a measure. *Journal of Personality Assessment, 65,* 322-342.

Tarnow, E. (2000). Self-destructive obedience in the airplane cockpit and the concept of obedience optimization. In T. Blass (Ed.), *Obedience to authority: Current perspectives on the Milgram paradigm* (pp. 111-123). Mahwah, NJ: Erlbaum.

Taylor, G. (1985). *Pride, shame, and guilt: Emotions of self-assessment.* Oxford: Clarendon.

Taylor, P. W. (1981). The ethics of respect for nature. *Environmental Ethics, 3,* 197-218.

Telfer, E. (1995). Self-respect. In R. S. Dillon (Ed.), *Dignity, character, and self-respect* (pp. 107-116). New York: Routledge. (Original work published 1968)

Tesser, A., & Shaffer, D. R. (1990). Attitudes and attitude change. *Annual Review of Psychology, 41,* 479-523.

Thomson, J. J. (1993). A defense of abortion. In J. Perry & M. Bratman (Eds.), *Introduction to philosophy: Classical and contemporary readings* (2nd. ed., pp. 497-507). New York: Oxford University Press. (Original work published 1971)

Timmons, M. (1985). Kant and the possibility of moral motivation. *The Southern Journal of Philosophy, 23,* 377-398.

Trafimow, D. (1997). The implications of success for hierarchically and partially restrictive ability dimensions. *Social Cognition, 4,* 312-326.

Urdang, L. (1992). *The Oxford thesaurus: American edition.* New York: Oxford University Press.

Van de Ven, P., Bornholt, L., & Bailey, M. (1996). Measuring cognitive, affective, and behavioral components of homophobic reaction. *Archives of Sexual Behavior, 25,* 155-179.

Velleman, J. D. (1999). Love as a moral emotion. *Ethics, 109,* 338-374.

Velleman, J. D. (2000). Well-being and time. In J. D. Velleman, *The possibility of practical reason* (pp. 56-84). Oxford: Clarendon. (Original work published 1991)

Vranas, P. B. M. (1995). *The practicality of morality: Facts and explanations*. Unpublished manuscript.

Vranas, P. B. M. (2000). *Evaluating inconsistent moral character: The asymptotic negativity effect.* Unpublished manuscript.

West, S. G., Gunn, S. P., & Chernicky, P. (1975). Ubiquitous Watergate: An attributional analysis. *Journal of Personality and Social Psychology, 32,* 55-65.

White, G., & Zimbardo, P. G. (1972). *The Stanford Prison Experiment* [Video]. Stanford University.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9,* 1-26.

Williams, B. (1976). Persons, character and morality. In A. O. Rorty (Ed.), *The identities of persons* (pp. 197-216). Berkeley, CA: University of California Press.

Wong, D. (1984). Taoism and the problem of equal respect. *Journal of Chinese Philosophy, 11,* 165-183.

Woodmansee, J. J., & Cook, S. W. (1967). Dimensions of verbal racial attitudes: Their identification and measurement. *Journal of Personality and Social Psychology, 7,* 240-250.

Wright, L. (1974). Emergency behavior. *Inquiry, 17,* 43-47.

Yakimovich, D., & Saltz, E. (1971). Helping behavior: The cry for help. *Psychonomic Science, 23,* 427-428.

Zanna, M. P., & Rempel, J. K. (1988). Attitudes: A new look at an old concept. In D. Bar-Tal & A. W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 315-334). Cambridge: Cambridge University Press.

Zimbardo, P. G. (1970). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In W. J. Arnold & D. Levine (Eds.), *Nebraska symposium on motivation, 1969* (pp. 237-307). Lincoln: University of Nebraska Press.

Zimbardo, P. G. (1973a). The psychological power and pathology of imprisonment. In O. Milton & R. G. Wahler (Eds.), *Behavior disorders: Perspectives and trends* (pp. 151-61). Philadelphia: Lippincott.

Zimbardo, P. G. (1973b). On the ethics of intervention in human psychological research: With special reference to the Stanford prison experiment. *Cognition, 2,* 243-256.

Zimbardo, P. G. (1975). Transforming experimental research into advocacy for social change. In M. Deutsch & H. A. Hornstein (Eds.), *Applying social psychology: Implications for research, practice, and training* (pp. 33-66). New York: Wiley.

Zimbardo, P. G., Haney, C., Banks, C., & Jaffe, D. (1973, April 8). The mind is a formidable jailer: A Pirandellian prison. *New York Times Magazine,* 38-60.

Zimbardo, P. G., Maslach, C., & Haney, C. (2000). Reflections on the Stanford Prison Experiment: Genesis, transformations, consequences. In T. Blass (Ed.), *Obedience to authority: Current perspectives on the Milgram paradigm* (pp. 193-237). Mahwah, NJ: Erlbaum.

Zimbardo, P. G., & White, G. (1972). The Stanford Prison Experiment: A simulation study of the psychology of imprisonment conducted August 1971 at Stanford University [Slide show]. Available: <http://zimbardo.com/prisonexp/>.

INDEX OF SUBJECTS AND AUTHOR NAMES

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z