The Indeterminacy Paradox: Character Evaluations and Human Psychology

PETER B. M. VRANAS
Iowa State University

Abstract

You may not know me well enough to evaluate me in terms of my moral character, but I take it you believe I *can* be evaluated: it sounds strange to say that I am *indeterminate*, neither good nor bad nor intermediate. Yet I argue that the claim that *most* people are indeterminate is the conclusion of a sound argument—the *indeterminacy paradox*—with two premises: (1) most people are *fragmented* (they would behave deplorably in many and admirably in many other situations); (2) fragmentation entails indeterminacy. I support (1) by examining psychological experiments in which most participants behave *deplorably* (e.g., by maltreating "prisoners" in a simulated prison) or *admirably* (e.g., by intervening in a simulated theft). I support (2) by arguing that, according to certain plausible conceptions, character evaluations presuppose behavioral consistency (*lack* of fragmentation). Possible reactions to the paradox include: (a) denying that the experiments are relevant to character; (b) upholding conceptions according to which character evaluations do not presuppose consistency; (c) granting that most people are indeterminate and explaining why it appears otherwise. I defend (c) against (a) and (b).

1. Introduction[1]

Imagine a person who regularly behaves admirably: he rescues people from burning buildings at considerable risk to his own life, he spends his summers working as a volunteer at a camp for blind children, and so on. Now suppose that the very same person regularly behaves deplorably as well: he swindles elderly people out of their life savings, he perjures himself for pay, and so on. How would you evaluate such a person in terms of his moral character?

A natural response is that such a person defies classification: he is in a sense both good and bad, and so is simpliciter neither. To coin a term, he is *indeterminate*: neither good nor bad nor intermediate. I will argue that this response is indeed a consequence of certain plausible conceptions of character evaluations which I call *consistency conceptions*: character evaluations presuppose behavioral consistency, so that an inconsistent or *fragmented* person, who behaves deplorably in many and admirably in many other situations, is indeterminate (details in §3). To say that a garden-variety person like you or me is indeterminate would sound strange, and to say that *most* people are indeterminate would be downright paradoxical, but to say that *fragmented* people are indeterminate presumably escapes paradox because fragmented people are exceptional. Are they really? Psychological experiments suggest otherwise.

In Milgram's obedience experiments, most participants were induced to administer powerful—in fact fictitious—electric shocks to a screaming confederate. In Zimbardo's prison experiment, most "guards" in a simulated prison maltreated the "prisoners". I will argue that these participants behaved deplorably: attempts to excuse them fail. Those who agree might be tempted to derive a bleak picture of human nature from these experiments, but a more balanced picture emerges when one also considers a set of experiments in which most participants behaved admirably: they helped an apparently electrocuted confederate at the risk of being electrocuted themselves, or they stopped a simulated theft. Taking both sets of experiments into account, I will conclude that *most* people are fragmented (details in §2). But then most people are also indeterminate, and we don't escape paradox after all. So let's call the following argument the *indeterminacy paradox*:[2]

(Q1) Most (i.e., the majority of) people are fragmented.
(Q2) Fragmentation entails indeterminacy.
Thus: (C) Most people are indeterminate.

One might react to this paradox by arguing that it commits the fallacy of equivocation: I motivated the claim that fragmentation entails indeterminacy by asking you to imagine a person who is fragmented in the sense that he *actually* behaves in *extremely* deplorable and admirable ways in many *real-life* situations, whereas the experiments suggest that most people are fragmented in the sense that they *would* behave in *mildly* deplorable and admirable ways in *artificial* (experimental) situations. Rest assured that I will address this objection at length: I will argue that, according to consistency conceptions of character evaluations, the kind of fragmentation whose prevalence is suggested by the experiments does suffice for indeterminacy (details in §3 and §4). Although I will argue that consistency conceptions are plausible, a second possible reaction to the paradox is to uphold one of two competing kinds of conceptions. According to *averaging conceptions*, character evaluations do not presuppose behavioral consistency because they function much like grade point averages: a student who gets many C's and many A's can still be a good or bad student if the A's far outweigh the C's or vice versa. According to *impurity conceptions*, goodness of character does presuppose consistency but badness does not: fragmented people are bad rather than indeterminate. I will respond by arguing that averaging and impurity conceptions are implausible (details in §3.4.2). My preferred reaction to the paradox is a third one: grant that most people are indeterminate, explain why it appears otherwise, and go on to conclude that character evaluations are unwarranted and should be replaced by *local* evaluations of people in light of their behavior in restricted ranges of situations (details in §5).

The indeterminacy paradox is motivated by a set of surprising psychological results (like Milgram's and Zimbardo's). 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 growing body of philosophical literature (e.g., Athanassoulis 2000; Bok 1996; Campbell 1999; Clarke 2003; DePaul 2000; Doris 1996, 1998, 2002; Flanagan 1991; Harman 1999, 2000, 2003; Kamtekar 2004; Kupperman 2001; Merritt 1999, 2000; Miller 2003; Pigden & Gillet 1996; Railton 1995; Solomon 2003; Sreenivasan 2002) 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). John Doris and Gilbert Harman, in particular, have proposed arguments in some respects similar to the indeterminacy paradox. The present work differs from those of Doris and Harman both methodologically (by deploying more formal arguments) and substantively. The main substantive differences are four. (1) Unlike Doris and Harman, I allow explicitly for the possibility that a *sizeable* minority of people are *not* fragmented. (2) Unlike Doris and Harman (who do not even introduce the notion of indeterminacy), I deny that fragmented people are *intermediate*: I insist that they are *indeterminate*. (3) Unlike Doris and Harman, I rely heavily on *counterfactual* behavior to argue that most people are fragmented and indeterminate. (4) The indeterminacy paradox is deductively valid, whereas Doris's (2002: 63) argument uses an inference to the best explanation. I expand on some of the above (and on some other) differences at various places later on.

In §2 I defend Q1. In §3 I defend Q2. In §4 I address an objection. In §5 I conclude by sketching my preferred reaction to the paradox.

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 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 (subjectively) *wrong* (in the sense of violating one's—all-things-considered—duty) and lacks an adequate excuse, 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). 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 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.[3]

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 the Appendix. In the remainder of the current section 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 two (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) and (2) Zimbardo's experiment (in which most "guards" in a simulated prison maltreated the "prisoners"). For each experiment I will argue that (I) most *participants* behaved deplorably and that (II) most *people* would behave as those participants who behaved deplorably did. Clearly, the existence of *two* 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) *obedience experiments*, 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 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 (Milgram 1974: 172) about 80% of the participants were nonsuspicious (so that about 64%—i.e., 80% of 80%—of the participants were nonsuspiciously obedient). 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. 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; contrast Ring, Wallston, & Corey 1970). 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 (like the attitude of those dentists who are blasé about 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"?[4] 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 (cf. O'Leary, Willis, & Tomich 1970: 91; contrast Mixon 1972: 150), whereas many participants displayed severe nervousness. 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.[5] Maybe the trust consisted in the belief that the experimenter had some (perhaps unfathomable) *scientifically* legitimate reason for conducting the experiment. 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* legitimate 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!) Maybe, however, the trust included in addition the belief that the experimenter had some such morally legitimate reason. I reply that such trust would at most *explain* but would again not morally *justify* 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.[6] 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. Doris 2002: 135; 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. Moreover, some actions are deplorable although almost everyone performs them: consider the inactivity of 38 witnesses to the Kitty Genovese murder (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 (e.g., Gilbert 1981). I reply that there was a natural disobedience point: 150 volts, when the learner withdrew his consent to continue (cf. 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. 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 (unbeknownst to the experimenter) 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 mitigating factors. But this concession is all I need: 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).[7] Second, high 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 et al. 1969, mentioned by Blass 2000: 48, 58–9), Spain (Miranda et al. 1981), UK (Burley & McGuinness 1977), and USA (e.g., Bok & Warren 1972; Costanzo 1976; Powers & Geen 1972; Rosenhan 1969: 141–3; cf. Shalala 1974).[8] 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. (In response to the claim that Milgram's experimental situation is too "artificial", in §4 I describe some more naturalistic related studies.)

2.2.2. The Stanford Prison Experiment

I turn now to the Stanford Prison Experiment, conducted by Zimbardo, Haney, Banks, and Jaffe (1973; cf. Zimbardo, Maslach, & Haney 2000). 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. The guards attended an orientation meeting in which they were told that 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 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, Banks, & Zimbardo 1973: 94, 95; cf. 1976: 173, 175.)
>
> [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. Even if the experimenters formulated only minimal guidelines, 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), 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; cf. Haney et al. 1973: 81, 94). 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; cf. Haney & Zimbardo 1977: 208; Zimbardo et al. 1973: 53; Zimbardo et al. 2000: 226). 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 necessitating 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" (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). 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" (Zimbardo et al. 1973: 48; cf. DeJong 1975: 1014) when his parole appeal was rejected? The blindness of some guards to the prisoners' suffering was presumably self-serving and does not adequately excuse those guards' behavior.[9]

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 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; cf. Evans 1980: 207; Zimbardo 1973a: 154; Zimbardo et al. 1973: 49; Zimbardo et al. 2000: 203).

(II) Two considerations support my claim that most *people* would behave much like those *participants* in the Stanford Prison Experiment (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,[10] 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 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. This completes my defense of the claim that (Q3) there are many situations in each of which most people (would) behave deplorably.[11]

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 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 electrocution experiments (in which most participants help an apparently electrocuted confederate at the risk of being electrocuted themselves) and (2) the theft experiments (in which most participants stop a simulated theft). For each experiment I will argue again that (I) most *participants* behaved admirably and that (II) most *people* would behave as those participants who behaved admirably did.

2.3.1. 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. Arguably there is a duty to assist a person in need when one has the power and the 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. 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; 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).[12] 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 (Wright 1974: 45). First, because such 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.

2.3.2. 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, 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 points to her suitcase on the table and 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). Did they behave admirably? First, it's hard to ascribe an ulterior motive to them.[13] (A speculation that they might have wanted to befriend the female "victim" won't do: in the beach experiment 9 out of 10 participants intervened when the victim was male and the thief female.) 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, but they could have intervened by just following the thief and *shouting* at him: the expected cost of a physical confrontation with the thief was high enough to make *stopping* him 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, 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 et al. (1980) found that all 10 participants who had agreed to watch a female victim'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; (iii) they also found in a replication that 5 out of 8 males and 7 out of 8 females intervened. Finally, in a large-scale experiment at a corridor of a university building, 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. All of the above investigators (as well as several others) 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.

This completes my defense of the claim that (Q4) there are many situations in each of which most people (would) behave admirably,[14] and thus of the claim that (Q1) most people are fragmented (although, as I said, the validity of the argument from Q3 and Q4 to Q1 is not obvious and is examined in the Appendix). Note that I did not defend the stronger (than Q1) claim that *everyone* is fragmented. Harman, however, seems to defend something like this stronger claim: he explicitly contrasts his view with claims which are only about *most* people (2000: 225). His reason is that "in Milgram (1963) *every* subject was willing to apply shocks of up to 300 volts" (Harman 2000: 225; cf. 1999: 322). I reply that, 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 which Harman refers "no vocal response or other sign of protest is heard from the learner until Shock Level 300 is reached" (Milgram 1963: 374); so arguably going up to 300 volts in the latter experiment was not deplorable. Maybe, however, Harman would accept—as Doris (1998: 524 n. 33, 2002: 65, 112, 193 n. 6) does—that situationist results are consistent with the *possibility* that a minority of people are not fragmented; so on a charitable reading Harman's point is that there is *no evidence* that this possibility is actual. Doris (1998: 511) suggests that this possible minority is small, but in the Appendix I derive an approximate lower bound on the percentage of fragmented people: 71%, as it turns out. But then the psychological results are consistent with the possibility that a *sizeable* minority (up to about 29%) 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.

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; cf. Doris 2002: 122), but why would I need to provide an answer? Maybe the question is 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.[15]

3. Fragmentation entails indeterminacy (Q2)

3.1. The concept of indeterminacy and three kinds of conceptions of character evaluations

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.[16] 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 character status: an indeterminate person is (e.g.) definitely not good.[17] 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.

Given the above understanding of indeterminacy, the claim that (Q2) fragmentation entails indeterminacy amounts to the claim that a person (i.e., *any*—actual or hypothetical—person) who behaves deplorably in many and admirably in many other situations is neither good nor bad nor intermediate. It can be seen that Q2 is equivalent to the conjunction of the following three claims (understood as quantified over *p*):[18]

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

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

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

(By "behaves" I mean "does or would behave" and by "many situations" I mean "an open list of actual or counterfactual situations".)

Whether one accepts Q2 depends on which conception of character evaluations one adopts. (1) Conceptions of character evaluations according to which Q2 is true 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[19] 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).[20] (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 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 (in the sense of requiring consistency neither of good nor of bad people)[21] and adopt a soft line on compensation. The following table summarizes the main characteristics of the above three kinds of conceptions.

	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)

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. I will give first an informal (§3.2) and a formal (§3.3) defense of consistency conceptions, and then I will address objections (§3.4).

3.2. An informal defense of consistency conceptions

Let us start with Q5 and 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. Nor does Q5 say that good people usually behave in the exact *same* way (cf. Doris 1996: 60–1, 2002: 176 n. 15): 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 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 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.

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".[22] This claim, like Q5, is about a "decent" person (επιεικήs), 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, 2002: 17–8). It seems then that I have plenty of company in finding Q5 intuitively appealing.

One can similarly defend Q6, given the symmetry between Q5 and Q6. (Later on I address a worry about Q6.) If Q5 and Q6 are true, then 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. To see why 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 capital punishment combines a multitude of factors, namely various considerations for and against. If I believe that neither group of considerations outweighs the other, then it's inaccurate to say that my attitude is between for and against: I rather have an *ambivalent* (cf. Priester & Petty 1996) attitude, which cannot be properly placed on a for/against scale.

In response one might argue that it would not be strange for me to choose the midpoint of a scale when questioned about my attitude towards capital punishment. I have two replies. First, the midpoint can be ambiguous: if it is designated as "neither for nor against" (or "neither good nor bad"), then it can correspond either to being intermediate or to being indeterminate. Second, we frequently choose midpoints not because they are appropriate, but because of situational pressures. Take a student paper which exhibits both outstanding 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), and is thus in *some* sense intermediate. I reply that this is not the *relevant* sense of 'intermediate'. To be intermediate in the sense of being *between* good and bad (and thus being *neither* good nor bad), a person must be between *every* good and bad person; this is not guaranteed by being between every *perfectly* good and bad person, 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.

This completes my informal defense of consistency conceptions. Next I offer a formal defense.

3.3. A formal defense of consistency conceptions

Are *every* two (possible) people comparable in terms of moral character? In other words, given two people, is it *always* the case that one of them is better (i.e., has a better character) than the other or they are equally good (or bad)?[23] Arguably not. If a first person is a model spouse and parent but a tyrannical employer whereas a second person is a tyrannical spouse and parent but a model employer, then it may well be the case that the two people are overall too different to be compared. Or at least that none of them is better than the other: I don't need to presuppose explicitly that they are not equally good (or bad). More specifically, what I do presuppose explicitly is the following claim:

> (Q8) If two people are such that the first behaves *much* better than the second in *many* situations and the second behaves *much* better than the first in *many* other situations, then none of them is (overall) better than the other.

If you accept Q8, then you are "trapped": Q8 turns out to entail Q2 (the claim that fragmentation entails indeterminacy). The proof has two steps: Q8 entails Q9 (stated below), and Q9 entails Q2.

> (Q9) For any fragmented person *f*, (i) there is a bad person *b* such that *f* is not better than *b* (i.e., it is not the case that *f* is better than *b*) and (ii) there is a good person *g* such that *g* is not better than *f*.

To see why Q8 entails Q9, take any fragmented person *f* and consider a bad person *b* who (a) behaves deplorably in every situation in which *f* behaves admirably and (b) behaves *neutrally* (i.e., neither blameworthily nor praiseworthily—hence neither deplorably nor admirably) in every other situation. (I am not taking a stand on whether *every* person who satisfies (a) and (b)—and thus behaves deplorably in an open list of situations but never behaves admirably or even praiseworthily—is bad; I am using only the weaker claim that at least *one* such person is bad.) Then *f* behaves much better than *b* in many situations (in which *f* behaves admirably and *b* deplorably), and *b* behaves much better than *f* in many other situations (in which *b* behaves neutrally and *f* deplorably). It follows, by Q8, that *f* is not better than *b*.[24] Similarly for Q9(ii).[25]

To see why Q9 entails Q2, take any fragmented person *f* and consider a bad person *b* and a good person *g* as in Q9. Given that (trivially) every good person is better than every bad person, *f* is not good: if *f* were good, then *f* would be better than every bad person, but *f* is not better than *b*. Similarly, *f* is not bad: if *f* were bad, then every good person would be better than *f*, but *g* is not better than *f*. Finally, given that (trivially) every intermediate person is better than every bad person, *f* is not intermediate either: if *f* were intermediate, then *f* would be better than every bad person, but *f* is not better than *b*. It follows that *f* is indeterminate.

This completes my formal argument for Q2. Next I address objections.

3.4. Objections to consistency conceptions

I will address first five objections to Q5 (§3.4.1) and then an objection to Q6 (§3.4.2).

3.4.1. 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 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 1970/1992; Doris 1998: 509–10, 2002: 16–7) 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:[26] 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.)

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 goodness 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*; 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*.[27] 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.

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 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.[28] 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).[29] I reply that there is no incompatibility because (as the 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.[30] 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]" (Smith 1991: 289; cf. Herman 1981: 368–9). Counterfactual behavior can be decisive for the latter kind of assessment (cf. Haybron 1999: 131) 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 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.[31])

3.4.2. An objection to Q6

Q6 says that a person who behaves admirably in many situations is not bad, regardless of how deplorably the person might also behave. But what about a person who, like Hitler, behaves *extremely* deplorably in many situations? Q6 does not have the consequence that Hitler was not bad, unless the (controversial) premise is accepted that Hitler behaved admirably in many situations. Consider, however, Schitler: a hypothetical dictator who orchestrates the murder of several million people but also (by assumption) behaves admirably in an open list of situations. Q6 does have the consequence that Schitler is not bad, and this consequence is wildly implausible—or so the objection goes. In reply let me clarify a bit the description of Schitler. I am not saying that Schitler is just nice to his friends and family: admirable behavior is by definition *highly* praiseworthy, like risking one's life to help an electrocuted stranger. Moreover, I am not talking about a few isolated instances of admirable behavior: an open list comprises by definition a *large* number of *multifarious* instances. But once these clarifications are really taken in, the claim that Schitler is not bad becomes to my mind much less implausible. (I am not saying that Schitler is good; so I do have a way to accommodate the claim that his admirable actions—and dispositions—do not *fully* compensate for his deplorable ones.) In response one might claim that Schitler's deplorable actions (far) outweigh his admirable ones. But even if this claim is true, what follows is at most that Schitler is *(much) more bad than good*, not that he is *bad*.[32] Schitler is a mixed bag; saying that he is bad does not do justice to the complexity of his character. Even if his deplorable actions *outweigh* his admirable ones, the former do not *obliterate* the latter. Schitler is in a way like an *idiot savant* who is a genius in some respects but an idiot in many others: it is inaccurate to say that such a person is an idiot.[33]

Despite the above considerations, I admit it is counterintuitive to deny that Schitler is bad. Impurity and averaging conceptions can grant that Schitler is bad, so they fare better than consistency conceptions in this respect. They fare worse, however, otherwise. Impurity conceptions have the counterintuitive implication that *mildly* deplorable behavior in an open list of situations guarantees badness even if accompanied by *extremely* admirable behavior in a *much larger* open list of other situations (*no* compensation is possible). Averaging conceptions have the counterintuitive implication that *extremely* deplorable behavior—like brutally raping and killing one's mother—in an open list of situations is compatible with goodness (*full* compensation is possible).[34] So none of these three kinds of conceptions of character evaluations is problem-free. To my mind, however, consistency conceptions strike the best balance between conflicting intuitions: they allow *some* but not *full* compensation (§3.1), so they can avoid the last two counterintuitive implications and they can grant that Schitler is not good. If so, then in the process of approaching reflective equilibrium a consistency conception should be adopted and the initially plausible claim that Schitler is bad should be revised.

I suspect that some people will not be convinced by my defense of the claim that Schitler is not bad. These people will insist that Schitler is bad, and they will thus reject consistency conceptions. They may grant my claim that consistency conceptions strike a better balance between conflicting intuitions than impurity and averaging conceptions do, but they may argue that the choice between these three kinds of conceptions amounts to a false trichotomy: other kinds of non-consistency conceptions also exist. In particular, they may want to defend a *weak impurity conception*, which claims that *extremely* deplorable behavior in an open list of situations guarantees badness. This claim still contradicts Q6 (the claim that admirable behavior in an open list of situations precludes badness), but is weaker than the claim (which impurity conceptions accept but weak impurity conceptions by definition reject) that *any* kind of deplorable behavior in an open list of situations guarantees badness. Weak impurity conceptions may be motivated by a wish to grant that garden-variety fragmented people like you or me need not be bad, coupled with a wish to insist that serial killers and perpetrators of genocide are always bad. Weak impurity conceptions avoid all three counterintuitive implications I mentioned above: (1) the implication of consistency conceptions that Schitler is not bad, (2) the implication of impurity conceptions that *mildly* deplorable behavior in an open list of situations guarantees badness, and (3) the implication of averaging conceptions that *extremely* deplorable behavior in an open list of situations is compatible with goodness.

Let me grant for the sake of argument that people who behave extremely deplorably in an open list of situations—call such people *Hitleresque*—are bad, as weak impurity conceptions claim. What about the remaining people, the non-Hitleresque ones? Weak impurity conceptions are compatible with the claim that (Q2′) *non-Hitleresque* people who are fragmented are indeterminate.[35] And this claim is entailed by the conjunction of Q5 and Q7 with the following modification of Q6: *non-Hitleresque* people who behave admirably in an open list of situations are not bad. Admittedly, on such a conception of character evaluations Q2 is still false: *some* possible fragmented people, namely the Hitleresque ones, are bad rather than indeterminate. But how many *actual* people are Hitleresque? For example, how many people (like Schitler) would orchestrate the murder of millions? Even if Q2 is false, it is arguably still the case that (Q2*) *almost all actual* fragmented people are indeterminate. If in addition, as I argued in §2, (Q1) most (actual) people are fragmented, it follows again that most people are indeterminate. We have thus a variant of the indeterminacy paradox which does not rely on consistency conceptions. This variant circumvents the objection that Schitler is bad. For the sake of simplicity in the remainder of the paper I talk only about the original paradox; those who reject Q2 can apply my claims mutatis mutandis to the variant of the paradox in which Q2* replaces Q2.

This completes for the moment my defense of the premises of the indeterminacy paradox. Next I address a more general objection to the paradox.

4. Experimental versus real-life fragmentation

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. Experimental fragmentation, however, does not suffice for indeterminacy, even if *real-life* 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 a reaction to the indeterminacy paradox 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. These considerations, however, are clearly inapplicable to my evidence for the prevalence of *admirable* behavior (§2.3): the electrocution and the theft 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 Milgram's and Zimbardo's experiments. I reply that there are also more naturalistic studies in which people behave deplorably. In an experiment by Hofling et al. (1966), nurses at two hospitals received a telephone call from an experimenter who identified himself as a physician and asked each nurse to administer what was obviously an 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 medication orders.[36] 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. More 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 (cf. Sreenivasan 2002: 59–61). 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 supposed 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 §3.4.1.) One might also argue that in extraordinary situations people may behave atypically, "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 (cf. Powell 1959: 496). (Moreover, can one behave ''out of character'' in an *open list* of situations?) More generally, the idea that only everyday-life behavior is relevant to character evaluations seems misguided: the relevance of behavior in a given situation to character evaluations depends (directly) not on how ordinary or extraordinary the situation is but on how trivial or significant the behavior is.[37] 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 don't need to use 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; it suffices instead to argue 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 *falsifies* the claim that the person is good (cf. Mook 1983: 383).

But 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?[38] One might object that in this case the deplorable behavior was *actual*, not counterfactual. In reply consider a questionnaire item from a study I conducted with introductory psychology students:

> 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 <u>very</u> 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, improbable, 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. Conclusion

A *paradox* is an apparently sound argument with an apparently unacceptable conclusion (see note 2). But an apparently sound argument may be *really* sound; an apparently unacceptable conclusion may be *true* after all. So in general a possible reaction to a paradox is to accept its conclusion, and this is indeed my preferred reaction to the indeterminacy paradox. The conclusion that most people are indeterminate is admittedly surprising, but this is small wonder given that the psychological results which lie at the root of the paradox are themselves surprising.

If most people are indeterminate, does it follow that our everyday practice of evaluating people in terms of their character is ungrounded? 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 do not have green eyes, we can reliably distinguish the minority of people who have green eyes from the majority who do not.) I believe that this possibility can be ruled out: an extensive literature in personality psychology suggests that from information on how a person behaves in certain situations we cannot confidently predict how the person would behave in other, dissimilar situations. Filling in the details is a long project which I undertake elsewhere. If that project is successful, then it does follow that character evaluations are epistemically unwarranted: we almost never have adequate evidence to evaluate with confidence particular people as good, bad, or intermediate. Call this conclusion the *epistemic thesis*.[39]

Suppose that the epistemic thesis is true. Suppose 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 most beneficial: how could you keep loving your mother if you regarded her as "fragmented", or 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) theists even if theism is epistemically unwarranted. In response—and following Doris (1996: 131, 1998: 507, 2002: 25)—I introduce my preferred alternative to character evaluations: *local* evaluations of people in light of their behavior in relatively restricted ranges of 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.[40] 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 the *comparative* claim that we have good pragmatic reason to *prefer* local to character evaluations (so that character evaluations are pragmatically unwarranted: they are not the most beneficial alternative). I wish to conclude this paper 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 than facing such situations with the misplaced confidence that I will overcome temptation (Doris 1996: 232–4, 1998: 515–7, 2002: 146–9). 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).

Giving a rigorous version of the above argument for the pragmatic thesis is a project for another occasion. Clearly, however, the prospects of that project in no way affect the success of my project in the present paper, which was to argue that the indeterminacy paradox is sound.

Appendix. The validity of the argument for the claim that (Q1) most people are fragmented

I said in §2 that the validity of the argument from Q3 and Q4 to Q1 is not obvious. Suppose, for example, that in each of three million situations 75%

of people behave deplorably. It doesn't follow that each 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. *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: $F_D/P \geq [(\pi_D/P) - (\sigma_D/S_D)]/[k_D - (\sigma_D/S_D)]$.*[41]

One can take the lower bound on F_D/P which is given by the above theorem 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. An estimate of π_D/P is 82%, namely the average of 64% and 100% (the approximate percentages of participants who behaved deplorably in Milgram's and Zimbardo's experiments). An estimate of k_D is 91%, namely the midpoint between 82% (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. It follows that an estimate of π_D/Pk_D, and thus an approximate lower bound on F_D/P, is $82\%/91\% \cong 90\%$.

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. A conservative estimate of π_A/P is 69%, namely the average of 62.5% and 76% (the approximate percentages of participants who behaved admirably in the electrocution and in Austin's theft experiments). An estimate of k_A (defined analogously to k_D) is 85%, the midpoint between 69% and one. By means of a theorem analogous to the above, it follows that an approximate lower bound on F_A/P is $69\%/85\% \cong 81\%$.

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 $F \geq F_D + F_A - P$,[42] so an approximate lower bound on F/P is $90\% + 81\% - 100\% = 71\%$: *most* people are fragmented.

Notes

I am very grateful to David Brink, Stephen Darwall, John Doris, Jack Foster, Allan Gibbard, Eric Lormand, Christian Miller, Peter Railton, Lyle Zynda, and some anonymous reviewers for extensive comments. I am also grateful to Elizabeth Anderson, Jami Anderson, John Devlin, David Hills, Nadeem Hussain, James Joyce, Frances Kamm, Richard Nisbett, Gerhard Nuffer, Andrew Robinson, Norbert Schwarz, Robert Stalnaker, Kevin Toh, David Velleman, Kendall Walton, and Michael Zimmerman for helpful discussion or comments. Thanks finally to many people, especially Gordon Belot, Karen Bennett, William Copeland, Fiona Cowie, Ann Cudd, Janine Diller, Craig Duncan, Samuel Floyd, Tamar Gendler, Gyl Gentzler, Peter Gibbard, John Haldane, Amy Kind, Celery Kovinsky, Rahul Kumar, Gerald Massey, George Mavrodes, Thomas Nagel, Gerald Postema, George Rainbolt, John Richardson, Samuel Ruhmkorff, Tony Smith, Peter Unger, and Jonathan Vogel for interesting questions. Material from this paper was presented at the 73rd Pacific APA (Berkeley, April 1999) and at the 97th Central APA (New Orleans, May 1999) meetings, as well as at the University of Michigan (Decision Consortium Conference, May 1999, and Department of Philosophy, September 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), Jesus College, Oxford (March 2001), and Peter Wall Institute for Advanced Studies, University of British Columbia (July 2001).

[1] This section presents the issues in a simplified and slightly imprecise way; rigor and comprehensiveness are found in later sections.

[2] I understand a *paradox* as an apparently sound argument with an apparently unacceptable conclusion (cf. Quine 1961/1976: 1; Sainsbury 1995: 1). The indeterminacy paradox (like Hempel's raven paradox but unlike Russell's set-theoretic paradox) is not an *antinomy*: it does not have a self-contradictory conclusion (cf. Quine 1961/1976: 5).

[3] I borrow the term 'fragmented' from Doris (1996: 134, 1998: 508, 2002: 25, 64), but I use it differently. As I read Doris, he calls a person 'fragmented' only if the person's behavior *actually* exhibits high variability. It follows that a person who is fragmented in my sense but always *actually* behaves admirably is not fragmented in Doris's sense. Conversely, a person who never *does or would* behave deplorably—and is thus not fragmented in my sense—but often behaves neither deplorably nor admirably and also often behaves *extremely* admirably exhibits high variability of behavior and can be fragmented in Doris's sense.

[4] 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 (i.e., *incontinently* rather than *viciously*). 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.

[5] 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: 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).

[6] 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.

[7] 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 volunteers have a higher need for social approval than nonvolunteers (Rosenthal & Rosnow 1975: 40–4), 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.

[8] 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".

[9] 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.

[10] 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). Also close to a replication comes the "BBC Prison Study", in which "the Guards failed to impose their authority and were eventually overcome by the Prisoners" (Reicher & Haslam 2004); but again, there were important procedural differences between this study and SPE. See also Orlando 1973 and Doyle 1975 for role-playing studies (not about prison environments) structurally similar to SPE.

[11] Further support for Q3 is provided by the *seizure experiments*, in which most participants fail to help a confederate who pretends to be the victim of an epileptic seizure and repeatedly asks for help (Darley & Latané 1968; cf. Evans 1980: 216–8; Hunt 1990: 132–5; Latané & Darley 1969: 261–5, 1970a: 22–5, 1970b: chap. 11, 1976: 14–7; for replications see Harris & Robinson 1973; Horowitz 1971; Schwartz & Clausen 1970; Schwartz & Gottlieb 1980).

[12] One might argue that touching the technician with one's hands was foolish and *therefore* not admirable. I object to the inference: even if entering a burning building to save a stranger is foolish, it may still be admirable.

[13] 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 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.

[14] Further support for Q4 is provided by the *rape experiments*, in which most participants try to stop a simulated rape (Anderson 1974, mentioned by Piliavin et al. 1981: 93, 133, 164, 172; Harari, Harari, & White 1985; Shotland & Straw 1976; contrast Shotland & Stebbins 1980).

[15] 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. One might argue, for example, that most people behave deplorably only in situations in which one can plausibly shift responsibility to someone else (e.g., an experimenter or an institution). But even if this is true, such situations are so numerous and multifarious that they include an open list.

[16] To say that some people are good, others are bad, and yet others are intermediate is not to deny that goodness and badness come in degrees. By definition, a person is intermediate exactly if she is better than every bad and worse than every good person.

[17] 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.

[18] Clearly, the conjunction of Q5, Q6, and Q7 entails Q2. To see why Q2 entails the conjunction, suppose that Q2 is true and consider a person p_1 who behaves deplorably in an open list of situations. Consider also 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 Q2) not good. But p_1 never behaves better than p_2 and is thus not good either. So Q2 entails Q5. Similarly, Q2 entails Q6. Finally, Q2 trivially entails Q7.

[19] 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.

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

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

[22] *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, 2002: 20–2; Flanagan 1991: 261–5, 282–3).

[23] Or they are "on a par" (Chang 1997: 4–5, 25–7; Qizilbash 2002: 141–6), "roughly comparable" (Parfit 1984: 431), "roughly equal" (Griffin 1986: 81).

[24] The gap between f's admirable and b's deplorable behavior in any of the former situations is larger than the gap between b's neutral and f's deplorable behavior in any of the latter situations, but it is a consequence of Q8 that this difference between the gaps does not matter for present purposes as long as both gaps are large enough. This consequence of Q8 might appear objectionable, but in fact it follows from the claim—which motivates Q8—that people who are too different are incomparable. (*Proof*. The last claim—which motivates Q8—entails that, if two people are too different to be comparable, then so are any two people who are even *more* different. Consider a person b' who is just like f except that b' behaves neutrally in every situation in which f behaves admirably. If b' and b are too different to be comparable, then so are f and b, who are even more different than b' and b.)

[25] How can a fragmented person be worse than *some* good persons (including the perfectly good ones, as I said in §3.2) but fail to be worse than *other* good persons (as Q8 entails)? To see how, take an analogy. Assign to each student paper a reasoning score and an originality score, each score being an integer from -5 to $+5$. Say that a paper is *good* exactly if both of its scores are at least $+3$, *bad* exactly if both of its scores are at most -3, and *fragmented* exactly if one of its scores is at least $+3$ and its other score is at most -3. Say also that a paper P_1 is *better* than a paper P_2 exactly if each score of P_1 is higher than the corresponding score of P_2. Then it can be seen that a paper is *intermediate* (in the sense of being better than every bad and worse than every good paper—see note 16) exactly if each of its scores is at least -2 and at most $+2$. It follows that every fragmented paper is *indeterminate*: neither good nor bad nor intermediate.

Moreover, a fragmented paper with reasoning score +4 and originality score −4 is worse than a good paper with both scores +5 but is not worse than a good paper with both scores +3.

[26] 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).

[27] 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.

[28] 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" (Smith 1991: 281–2; cf. Weiner 2003: 177–8). It is a matter of debate whether Kant (*Groundwork* 4: 398) adopts such a model (cf. Benson 1987; Henson 1979; Herman 1981).

[29] 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 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.

[30] 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). On the other hand, what Nagel calls "constitutive luck" can be relevant to the character one has and thus also to character evaluations.

[31] 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. (So, contrary to what Doris (2002: 118, 208 n. 32) suggests, I do *not* claim that "the indeterminacy regarding counterfactual situations is precisely what undermines person evaluation".)

[32] But if Schitler is more bad than good, how can he be indeterminate? Recall that the claim that a person is indeterminate is the claim that the person has no status on the good/intermediate/bad scale; this is compatible with the claim that the person has status on some "mixed" (multidimensional) scale.

[33] One can similarly reply to a similar objection against Q8: the objection that Q8 has the implausible consequence that Schitler is not worse than a person who always behaves neutrally, and is even not worse than *some* good person (see §3.3). But then, one might ask, what reason do we have to avoid being like Schitler? We have at least two reasons. (1) Schitler is not better than *some* bad person (see §3.3), and we have reason to be better than *every* bad person. (2) Schitler behaves extremely deplorably, and we have reason to avoid such behavior.

[34] One might deny that averaging conceptions have this implication: arguably some deplorable actions are so extreme that no admirable action is equally extreme. (As an analogy, no good grade is as extreme as F: arguably A+ is as extreme as C−, not F.) I reply that a sufficiently large number of less extreme admirable actions can still outweigh a smaller number of more extreme deplorable ones. (To pursue the analogy, a sufficiently large number of A+'s will result in a high GPA even in the presence of some F's.)

[35] One might point out that weak impurity conceptions are *also* compatible with the claim that non-Hitleresque people are never indeterminate, but are rather good, intermediate, or bad

depending on the balance between their praiseworthy and their blameworthy actions. That claim, however, has the counterintuitive implication that Q5 is false. (Q5 has the consequence that *mildly* deplorable behavior in an open list of situations *precludes goodness*. Asserting—as I do—the plausibility of this consequence is compatible with denying—as I do—the plausibility of the claim that mildly deplorable behavior in an open list of situations *guarantees badness*.)

[36] 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).

[37] This is not to deny that how ordinary or extraordinary a situation is can affect how trivial or significant a behavior in that situation is and can thus *indirectly* affect the relevance of the behavior to character evaluations.

[38] But why didn't they also break off their friendships with some of the *prisoners*? One might argue that this was either (1) because they failed to realize that some of the prisoners would also have behaved sadistically if they had role-played guards, or (2) because they cared about the *actual* sadistic behavior of the guards, not about the *counterfactual* sadistic behavior of the prisoners. Arguably, however, it was instead (3) because they had no way of knowing *which* of the prisoners would have behaved sadistically.

[39] Doris agrees that evaluating people as good or bad is epistemically unwarranted (1998: 514, 2002: 115; contrast 1996: 127), but apparently he does not endorse my epistemic thesis because he rejects my claim that fragmentation entails indeterminacy (cf. 2002: 115): apparently he would confidently evaluate many people as intermediate.

[40] 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.

[41] *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 $\Sigma_p\ S_{Dp} = \Sigma_{p\varepsilon \boldsymbol{F}}\ S_{Dp} + \Sigma_{p\varepsilon \boldsymbol{F}'}\ S_{Dp} \leq k_D S_D F_D + \sigma_D(P - F_D)$ (3). Combining (1), (2), and (3), we get: $S_D \pi_D \leq k_D S_D F_D + \sigma_D(P - F_D)$, which is equivalent to $F_D \geq (S_D \pi_D - \sigma_D P)/(k_D S_D - \sigma_D)$, from which the theorem immediately follows. □ (Applied to the numerical example I gave in the text, the theorem gives: $F_D/P \geq (75\% - 1/3)/(k_D - 1/3) \geq 62.5\%$ because $k_D \leq 1$.)

[42] *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 $F \geq F_D + F_A - P$. □

References

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, IN: Hackett.

Athanassoulis, N. (2000) A response to Harman: Virtue ethics and character traits *Proceedings of the Aristotelian Society* 100: 215–221.

Austin, W. (1979) Sex differences in bystander intervention in a theft *Journal of Personality and Social Psychology* 37: 2110–2120.

Badhwar, N. K. (1996) The limited unity of virtue *Noûs* 30: 306–329.

Banuazizi, A., Movahedi, S. (1975) Interpersonal dynamics in a simulated prison: A methodological analysis *American Psychologist* 30: 152–160.

Benson, P. (1987) Moral worth *Philosophical Studies* 51: 365–382.

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 *Advances in Experimental Social Psychology* 25: 277–329.

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.

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.

Brandt, R. B. (1992) 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)

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.

Campbell, J. (1999) Can philosophical accounts of altruism accommodate experimental data on helping behaviour? *Australasian Journal of Philosophy* 77: 26–45.

Chang, Ruth. (Ed.). (1997) *Incommensurability, incomparability, and practical reason.* Cambridge, MA: Harvard University Press.

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.

Clarke, S. (2003) *Courage under fire: Is there really a fundamental attribution error?* Unpublished manuscript.

Costanzo, 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.

Darley, J. M. (1995) Constructive and destructive obedience: A taxonomy of principal-agent relationships *Journal of Social Issues* 51: 125–154.

Darley, J. M., Latané, B. (1968) Bystander intervention in emergencies: Diffusion of responsibility *Journal of Personality and Social Psychology* 8: 377–383.

DeJong, W. (1975) Another look at Banuazizi and Movahedi's analysis of the Stanford Prison Experiment *American Psychologist* 30: 1013–1015.

DePaul, M. (2000) Character traits, virtues, and vices: Are there none? In B. Elevitch (Ed.), *The Proceedings of the Twentieth World Congress of Philosophy: Vol. 9. Philosophy of mind* (pp. 141–157). Bowling Green, OH: Philosophy Documentation Center.

Doris, J. M. (1996) *People like us: Morality, psychology, and the fragmentation of character.* Doctoral dissertation, The University of Michigan.

Doris, J. M. (1998) Persons, situations, and virtue ethics *Noûs* 32: 504–530.

Doris, J. M. (2002) *Lack of character: Personality and moral behavior*. New York: Cambridge University Press.

Doyle, C. L. (1975) Interpersonal dynamics in role playing *American Psychologist* 30: 1011–1013.

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.

Evans, R. I. (Ed.). (1980) *The making of social psychology: Discussions with creative contributors*. New York: Gardner Press.

Faber, N. (1971, October 15) 'I almost considered the prisoners as cattle' *Life* 71: 82–3.

Flanagan, O. (1991) *Varieties of moral personality: Ethics and psychological realism*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1983) *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

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.

Griffin, J. (1986) *Well-being: Its meaning, measurement and moral importance*. Oxford, England: Clarendon Press.

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., Banks, W. C., Zimbardo, P. G. (1973) Interpersonal dynamics in a simulated prison *International Journal of Criminology and Penology* 1: 69–97.

Haney, C., Banks, W. 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: Psychological and legal issues* (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 *The 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.

Harman, G. (2003) No character or personality *Business Ethics Quarterly* 13: 87–94.

Harré, R. (1979) *Social being: A theory for social psychology*. Oxford, England: Basil Blackwell.

Harris, V. A., Robinson, C. E. (1973) Bystander intervention: Group size and victim status *Bulletin of the Psychonomic Society* 2: 8–10.

Haybron, D. M. (1999) Evil characters *American Philosophical Quarterly* 36: 131–148.

Henson, R. G. (1979) What Kant might have said: Moral worth and the overdetermination of dutiful action *The Philosophical Review* 88: 39–54.

Herman, B. (1981) On the value of acting from the motive of duty *The Philosophical Review* 90: 359–382.

Hofling, C. K., Brotzman, E., Darlymple, S., Graves, N., Pierce, C. M. (1966) An experimental study in nurse-physician relationships *The Journal of Nervous and Mental Disease* 143: 171–180.

Horowitz, I. A. (1971) The effect of group norms on bystander intervention *The Journal of Social Psychology* 83: 265–273.

Hunt, M. (1990) *The compassionate beast: What science is discovering about the humane side of humankind*. New York: William Morrow.

Ingram, P. (1979) Deception, obedience and authority *Philosophy* 54: 529–533.

Irwin, T. H. (Trans.). (1985) *Aristotle: Nicomachean Ethics*. Indianapolis, IN: Hackett.

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, England: Clarendon Press.

Kamtekar, R. (2004) Situationism and virtue ethics on the content of our character *Ethics* 114: 458–491.

Kant, I. (1964) *Groundwork of the metaphysic of morals* (H. J. Paton, Trans.). New York: Harper & Row. (Original work published 1785)

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.

Kupperman, J. J. (1991) *Character*. New York: Oxford University Press.

Kupperman, J. J. (2001) The indispensability of character *Philosophy* 76: 239–250.

Landis, C. (1924) Studies of emotional reactions: II. General behavior and facial expression *The 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: Social psychological studies of some antecedents and consequences* (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.

Lifton, R. J. (1986) *The Nazi doctors: Medical killing and the psychology of genocide*. New York: Basic Books.

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.

Mandelbaum, M. (1955) *The phenomenology of moral experience*. Glencoe, IL: Free Press.

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* 15: 239–245.

Meeus, W. H. J., Raaijmakers, Q. A. W. (1995) Obedience in modern society: The Utrecht studies *Journal of Social Issues* 51: 155–175.

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.

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. (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*. Westport, CT: Praeger.

Miller, A. G. (1995) Constructions of the obedience experiments: A focus upon domains of relevance *Journal of Social Issues* 51: 33–53.

Miller, C. (2003) Social psychology and virtue ethics *The Journal of Ethics* 7: 365–392.

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.

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 normality 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.

Moriarty, T. (1975) Crime, commitment, and the responsive bystander: Two field experiments *Journal of Personality and Social Psychology* 31: 370–376.

Musen, K., Zimbardo, P. G. (1992) *Quiet rage: The Stanford prison study* [Video]. Stanford University.

Nagel, T. (1979) Moral luck. In T. Nagel, *Mortal questions* (pp. 24–38). New York: Cambridge University Press. (Original work published 1976)

O'Leary, C. J., Willis, F. N., Tomich, E. (1970) Conformity under deceptive and non-deceptive techniques *The Sociological Quarterly* 11: 87–93.

Orlando, N. J. (1973) The mock ward: A study in simulation. In O. Milton & R. G. Wahler (Eds.), *Behavior disorders: Perspectives and trends* (3rd ed., pp. 162–170). Philadelphia: Lippincott.

Orne, M. T., Holland, C. H. (1968) On the ecological validity of laboratory deceptions *International Journal of Psychiatry* 6: 282–293.

Parfit, D. (1984) *Reasons and persons*. Oxford, England: Clarendon Press.

Patten, S. C. (1977a) The case that Milgram makes *The Philosophical Review* 86: 350–364.

Patten, S. C. (1977b) Milgram's shocking experiments *Philosophy* 52: 425–440.

Pigden, C. R., Gillet, G. R. (1996) Milgram, method and morality *Journal of Applied Philosophy* 13: 233–250.

Piliavin, J. A., Dovidio, J. F., Gaertner, S. L., Clark, R. D., III. (1981) *Emergency intervention*. New York: Academic Press.

Powell, B. (1959) Uncharacteristic actions *Mind* 68: 492–509.

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.

Priester, J. R., Petty, R. E. (1996) The gradual threshold model of ambivalence: Relating the positive and negative bases of attitudes to subjective ambivalence *Journal of Personality and Social Psychology* 71: 431–449.

Qizilbash, M. (2002) Rationality, comparability and maximization *Economics and Philosophy* 18: 141–156.

Quine, W. V. (1976) The ways of paradox. In W. V. Quine, *The ways of paradox and other essays* (pp. 1–18). Cambridge, MA: Harvard University Press. (Original work published 1961)

Railton, P. (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.

Reicher, S., Haslam, S. A. (2004) *Rethinking the psychology of tyranny*. Unpublished manuscript.

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.

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, M. (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.

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., Rosnow, R. L. (1975) *The volunteer subject*. New York: Wiley.

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.

Sabini, J., Silver, M. (1982) *Moralities of everyday life*. New York: Oxford University Press.

Sainsbury, R. M. (1995) *Paradoxes* (2nd ed.). New York: Cambridge University Press.

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. (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 *The Journal of Social Psychology* 110: 183–192.

Shaffer, D. R., Rogel, M., Hendrick, C. (1975) Intervention in the library: The effect of increased responsibility on bystanders' 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., 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.

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*, pp. 165–166.

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.

Smith, H. M. (1991) Varieties of moral worth and moral credit *Ethics* 101: 279–303.

Smith, P. B., Bond, M. H. (1993) *Social psychology across cultures: Analysis and perspectives.* New York: Simon & Schuster.

Solomon, R. C. (2003) Victims of circumstances? A defense of virtue ethics in business *Business Ethics Quarterly* 13: 43–62.

Sreenivasan, G. (2002) Errors about errors: Virtue theory and trait attribution *Mind* 111: 47–68.

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.

Weiner, B. (2003) A naïve psychologist examines bad luck and the concept of responsibility *The Monist* 86: 164–180.

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.

Wright, L. (1974) Emergency behavior *Inquiry: An Interdisciplinary Journal of Philosophy and the Social Sciences* 17: 43–47.

Zimbardo, P. G. (1973a) The psychological power and pathology of imprisonment. In O. Milton & R. G. Wahler (Eds.), *Behavior disorders: Perspectives and trends* (3rd ed., 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, W. C., Jaffe, D. (1973, April 8) The mind is a formidable jailer: A Pirandellian prison *The New York Times Magazine*, pp. 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 [On-line slide show]. Available: <http://www.prisonexp.org/>.