PDF version of the entry
Weakness of Will
http://plato.stanford.edu/archives/spr2009/entries/weakness-will/

from the Spring 2009 Edition of the

Stanford Encyclopedia of Philosophy

Library of Congress Catalog Data
ISSN: 1095-5054

Weakness of Will

First published Wed May 14, 2008

1. Julie chose *b* over *a*, even though she knew *b* was more expensive than *a*.

There is nothing puzzling about Julie's choice. Perhaps Julie was choosing among vacation options, and *b* was a week's vacation in Paris, while *a* was a week's vacation in Peoria. In any event, Julie evidently took the overall merits of *b* to outweigh those of *a*, even if *b* was inferior from a financial standpoint.

2. Jimmy opted for *d* over *c*, despite his judging *c* to be a healthier choice than *d*.

Again, we find Jimmy's decision unremarkable. Perhaps *c* and *d* were competing dessert options: *c*, let us suppose, was a dish of dry Wheaties, rich in fiber and whole grain, whereas *d* was a gossamer-light yet oh-so-rich Valrhona chocolate mousse. Jimmy obviously (and reasonably!) assessed *d* as the better dessert option all things considered, even though he knew *d* was less good for his health than *c* would be. Nothing puzzling about that.

3. Joseph did *f* rather than *e*, even though he was convinced that *e* was the better thing to do all things considered.

Here, by contrast, we have a genuinely puzzling case, one we cannot make sense of in the same way. Why would Joseph do *f* when he assessed *e* as the superior course of action *all things considered*? Joseph's choice sounds so inexplicable that we might even query whether the case has been accurately described. If Joseph is really freely choosing *f* over *e*, we might think it questionable that he does genuinely assess *e* as better all

things considered. Perhaps he actually takes *f* to be superior (for him, under the circumstances), although he thinks most people would opt for *e* or would say *e* was a better choice.

Our divergent reactions to these three examples point to something distinctive about the judgment that one course of action is better than another. (Better *overall*, or better all things considered, that is—not simply better in some respect.) Such judgments appear to enjoy a special connection to the agent's actions which other judgments do not possess. We are puzzled by Joseph's choice precisely because we expect people's actions—at least when freely undertaken—to reflect their overall assessment of the merits of the alternative courses of action before them. We expect their actions, in other words, to reflect that special judgment. And Joseph's—at least as reported above—doesn't. When judgment and action are said to have diverged in this way, we are often sceptical: we question whether the agent really held the course of action not taken to be better. And even when we accept the description of the case, we find such action somehow puzzling, defective, or dubiously intelligible, in a way that action contrary to one's judgments of financial wisdom (for example) is not. We can conclude that the particular judgment contrary to which Joseph acts—the judgment that one course of action is better than another—has what we can vaguely term a special character, in comparison with other judgments such as that one course of action is healthier than another.

Let us give a name to the assessment of his options contrary to which Joseph acts. Let us call his judgment that *e* is a better thing to do all things considered Joseph's *better judgment*. ("Better judgment" does not mean "superior judgment"; it simply means a judgment as to which option is overall better.) Joseph, then, appears to have acted, freely and intentionally, contrary to his better judgment. And this is precisely the phenomenon the philosophical tradition calls "weakness of will."[1]

Philosophers have been perplexed by or dubious about such action for a very long time.[2] Indeed, Plato's Socrates famously denied its possibility in the *Protagoras*. "No one," he declared, "who either knows or believes that there is another possible course of action, better than the one he is following, will ever continue on his present course" (*Protagoras* 358b-c). And philosophers have been wrestling with the issue ever since. It is not surprising that weakness of will has such a long and distinguished pedigree as a topic of philosophical discussion: it is both an intrinsically interesting phenomenon and a topic rich in implications for our broader theories of action, practical reasoning, rationality, evaluative judgment, and the interrelations among these.

---

1. Hare on the Impossibility of Weakness of Will

Let us commence our examination of contemporary discussions of this issue in appropriately Socratic vein, with an account that gives expression to and builds on many of the intuitions that lead us to be sceptical about reports like (3) above. For the moral philosopher R. M. Hare—as for Socrates—it is impossible for a person to do one thing if he genuinely and in the fullest sense holds that he ought instead to do something else. (If, that is—to echo the earlier quote from Socrates—he "believes that there is

another possible course of action, better than the one he is following.") This certainly seems to constitute a denial of the possibility of akratic or weak-willed action. In Hare's case it is a consequence of the general account of the nature of evaluative judgments which he defends (Hare 1952; see also Hare 1963).

Hare is much impressed by what we vaguely referred to above as the "special character" of evaluative judgments: judgments, that is, such as that one course of action is *better* than another, or that one *ought* to do a certain thing. Such evaluative judgments seem to have properties that differentiate them from merely "descriptive" judgments such as that one thing is more expensive than another, or rounder than another (Hare 1952, p. 111). Evaluative judgments seem, in particular, to bear a special connection to *action* which no purely descriptive judgment possesses. Hare's analysis, then, takes off from something like the data we rehearsed earlier. Hare goes on to develop these data in the following way. He begins by identifying, as the fundamental distinctive feature of evaluative judgments—that which lends them a special character—that evaluative judgments are intended to *guide conduct*. (See, e.g., Hare 1952, p. 1; p. 29; p. 46; p. 125; p. 127; p. 142, pp. 171–2; Hare 1963, p. 67; p. 70.) The special function of evaluative judgments is to be action-guiding: that is, if you will, what evaluative judgments are *for*. Hare then puts a more precise gloss on what it is for a judgment to "guide conduct": an action-guiding judgment is one which entails an answer to the practical question "What shall I do?" (Hare 1952, p. 29; see Hare 1963, p. 54 for the terminology "practical question").[3] What is it that an action-guiding judgment must entail? That is, what constitutes an answer to the question "What shall I do?" Hare holds that no (descriptive) statement can constitute an answer to such a question (Hare 1952, p. 46). Rather, such a question is answered by a first-person *command* or *imperative* (Hare 1952, p. 79), which could be verbally expressed as "Let me do *a*" (Hare 1963, p. 55).

To recap the argument thus far: it is the function of evaluative judgments like "I ought to do *a*" to guide conduct. Guiding conduct means entailing an answer to the question "What shall I do?" An answer to that question will take the form "Let me do *a*," where this is a first-person command or imperative. Therefore evaluative judgments entail such first-person imperatives (Hare 1952, p. 192). Now in general, if judgment J_1 entails judgment J_2, then assenting to J_1 must involve assenting to J_2: someone who professed to assent to J_1 but who disclaimed J_2 would be held not to have spoken correctly when he claimed to assent to J_1 (Hare 1952, p. 172). So assenting to an evaluative judgment like "I ought to do *a*" involves assenting to the first-person command "Let me do *a*" (Hare 1952, pp. 168–9). We should inquire, then, what exactly is involved in sincerely assenting to a first-person command or imperative of this type. Just as sincere assent to a statement involves *believing* that statement, sincere assent to an imperative addressed to ourselves involves *doing* the thing in question:

> It is a tautology to say that we cannot sincerely assent to a … command addressed to ourselves, and *at the same time* not perform it, if now is the occasion for performing it and it is in our (physical and psychological) power to do so (Hare 1952, p. 20).

So: provided it is within my power to do *a* now, if I do not do *a* now it follows that I do not genuinely judge that I ought to do *a* now. Thus, as Hare states at the very opening of his book, a person's evaluative judgments are infallibly revealed by his actions and choices:

> If we were to ask of a person 'What are his moral principles?' the way in which we could be most sure of a true answer would be by studying what he *did*…. It would be when … he was faced with choices or decisions between alternative courses of action, between alternative answers to the question 'What shall I do?',

> that he would reveal in what principles of conduct he really believed (Hare 1952, p. 1).

Note that Hare is not simply saying that a person's actions are the most reliable source of evidence as to his evaluative judgments, or that if a person did *b* the most likely hypothesis is that he judged *b* to be the best thing to do. Hare is saying, rather, that it *follows* from a person's having done *b* that he judged *b* best from among the options open to him at the time. On this view, then, akratic or weak-willed actions as we have understood them are impossible. There could not be a case in which someone genuinely and in the fullest sense held that he ought to do *a* now (where *a* was within his power) and yet did *b*. On Hare's view, "it becomes analytic to say that everyone always does what he thinks he ought to [if physically and psychologically able]" (Hare 1952, p. 169).

But *does* everyone always do what he thinks he ought to, when he is physically and psychologically able? It may seem that this is simply not always the case (even if it is *usually* the case). Have you, dear reader, *never* failed to get up off the couch and turn off the TV when you judged it was really time to start grading those papers? Have you *never* had one or two more drinks than you thought best on balance? Have you *never* deliberately pursued a sexual liaison which you viewed as an overall bad idea? In short, have you *never* acted in a way which departed from your overall evaluation of your options? If so, let me be the first to congratulate you on your fortitude. While weak-willed action does seem somehow puzzling, or defective in some important way, *it does nonetheless seem to happen*.

For Hare, however, any apparent case of *akrasia* must in fact be one in which the agent is actually *unable* to do *a*, or one in which the agent does not genuinely evaluate *a* as better—even if he says he does.[4] As an example of the first kind of case Hare cites Medea (Hare 1963, pp. 78–9),

who (he contends) is powerless, literally helpless, in the face of the strong emotions and desires roiling her: she is truly *unable* (psychologically) to resist the temptations besieging her. A typical example of the second kind of case, on the other hand, would be one in which the agent is actually using the evaluative term "good" or "ought" only in what Hare calls an "inverted-commas" sense (Hare 1952, p. 120; pp. 124–6; pp. 164–5; pp. 167–171). In such cases, when the agent says (while doing *b*) "I know I really ought to do *a*," he means only that most people—or, at any rate, the people whose opinions on such matters are generally regarded as authoritative—would say he ought to do *a*. As Hare notes (Hare 1952, p. 124), to believe this is not to make an evaluative judgment oneself; rather, it is to allude to the value-judgments of other people. Such an agent does not himself assess the course of action he fails to follow as better than the one he selects, even if other people would.

No doubt there are cases of the two types Hare describes; but they do not seem to exhaust the field. We can grant that there is the odd murderer, overcome by irresistible homicidal urges but horrified at what she is doing. But surely not every case that we might be tempted to describe as one of acting contrary to one's better judgment involves irresistible psychic forces. Consider, for example, the following case memorably put by J. L. Austin:

> I am very partial to ice cream, and a bombe is served divided into segments corresponding one to one with the persons at High Table: I am tempted to help myself to two segments and do so, thus succumbing to temptation and even conceivably (but why necessarily?) going against my principles. But do I lose control of myself? Do I raven, do I snatch the morsels from the dish and wolf them down, impervious to the consternation of my colleagues? Not a bit of it. We often succumb to temptation with calm and even with finesse (Austin 1956/7, p. 198).

(I might add that it also seems doubtful that irresistible psychic forces kept you on the couch watching TV while those papers were waiting.) As for the "inverted-commas" case, this too surely happens: people do sometimes pay lip service to conventional standards which they themselves do not really accept. But again, it seems highly doubtful that this is true of all seeming cases of weak-willed action. It seems depressingly possible to select and implement one course of action while *genuinely believing* that it is an overall worse choice than some other option open to you.

Has something gone wrong? We started with the unexceptionable-sounding thought that moral and evaluative judgments are intended to guide conduct; we arrived at a blanket denial of the possibility of akratic action which fits ill with observed facts. But if we are disinclined to follow Hare this far we should ask what the alternative is, for it may be even worse. For Hare, the answer is clear: our only other option is to repudiate the idea that moral and other evaluative judgments have a special character or nature, namely that of being action-guiding. For we should recall that Hare presents all his subsequent conclusions as simply following, through a series of steps, from that initial thought. "The reason why actions are in a peculiar way revelatory of moral principles is that the function of moral principles is to guide conduct," Hare continues in the passage quoted earlier (Hare 1952, p. 1). For Hare, then, the only way to escape his "Socratic" conclusion about weakness of will would be to give up the idea that evaluative judgments are intended to guide conduct, or to "have [a] bearing upon our actions" (Hare 1963, p. 169; see also Hare 1952, p. 46; p. 143; p. 163; pp. 171–2; and Hare 1963, p. 70; p. 99).

The choices before us, then, as presented by Hare, are Hare's own view, or one which assigns no distinctive role in action or practical thought to evaluative judgments, treating them as just like any other judgment. We might call the first of these an extreme version of (judgment) *internalism*.

(I use this polysemous label to refer, here, to the idea that certain judgments have an internal or necessary connection to motivation and to action.) By extension, we might usefully follow Michael Bratman in calling the second type of view "extreme externalism" (Bratman 1979, pp. 158–9).

Extreme externalism also seems unsatisfactory, however. First, it seems unable to explain why there should be anything perplexing or problematical about action contrary to one's better judgment, why there should be any philosophical problem about its possibility or its analysis. On this kind of view, it seems, Joseph's choice ((3) above) should strike us as no more puzzling than Julie's or Jimmy's ((1) or (2)). As Hare puts it:

> On the view that we are considering, there is nothing odder about thinking something the best thing to do in the circumstances, but not doing it, than there is about thinking a stone the roundest stone in the vicinity and not picking it up, but picking up some other stone instead.… There will be nothing that requires explanation if I choose to do what I think to be, say, the worst possible thing to do and leave undone what I think the best thing to do (Hare 1963, pp. 68–9).

But our reactions to (1), (2), and (3) show that we *do* think there is something peculiar about action contrary to one's better judgment which renders such action hard to understand, or perhaps even impossible. An extreme externalist view thus seems to mischaracterize the status of akratic actions.

Perhaps even more importantly, however, extreme externalism has dramatic implications for our understanding of intentional action in general—not just weak-willed action. For such a view implies that

> deliberation about what it would be best to do has no closer relation to practical reasoning than, say, deliberation about what it would be chic to do. If one happens to care about what it would be chic to do, then a consideration of this matter may play an important role in one's practical reasoning. If one does not care, it will be irrelevant. The case is the same with reasoning about what it would be best to do (Bratman 1979, p. 158).

To adopt a general doctrine of this sort seems an awfully precipitous response to the possibility of *akrasia*. For it seems extremely plausible to assign to our overall evaluations of our options an important role in our choices. Man is a rational animal, the saying goes; that is—to offer one gloss on this idea—we act on reasons, and in the light of our assessments of the overall balance of reasons. When we engage in deliberation or reasoning about what to do, we often proceed by thinking about the reasons which favor our various options, and then bringing these together into an overall assessment which is, precisely, intended to guide our choice.

Or, as Bratman puts it, we very often reason about what it is *best* to do as a way of settling the question of what *to* do. (He calls this "evaluative practical reasoning": Bratman 1979, p. 156.) "One's evaluations [thus] play a crucial role in the reasoning underlying full-blown action," Bratman holds (p. 170), and to be forced to deny this would be in his view "too high a price to pay" (p. 159). As Alfred Mele similarly puts it, "there is a real danger that in attempting to make causal and conceptual space for full-fledged akratic action one might commit oneself to the rejection of genuine ties between evaluative judgment and action" (1991, p. 34). But that would be to throw the baby out with the bathwater. If we want to resist Hare's conclusions, we must do so in a way which steers clear of the danger to which Mele alerts us. We must navigate between the Scylla of extreme internalism and the Charybdis of extreme externalism.

2. Davidson on the Possibility of Weakness of Will

This is just what Donald Davidson set out to do in a rich, elegant, and incisive paper published in 1970 which has had a towering influence on the subsequent literature. Davidson's treatment aims to vindicate the possibility of weakness of will; to offer a novel analysis of its nature; to clarify its status as a marginal, somehow defective instance of agency which we rightly find dubiously intelligible; and to do all this within the contours of a general view of practical reasoning and intentional action which assigns a central and special role to our evaluative judgments. Let us see how he proposes to do these things.

First, Davidson offers the following general characterization of weak-willed or incontinent action:[5]

> In doing *b* an agent acts incontinently if and only if: (a) the agent does *b* intentionally; (b) the agent believes there is an alternative action *a* open to him; and (c) the agent judges that, all things considered, it would be better to do *a* than to do *b*.

We initially described weak-willed action as free, intentional action contrary to the agent's better judgment; it may be useful to see how Davidson's more precise definition matches up with that initial characterization. Davidson's condition (a) requires that the action in question be intentional.[6] Condition (b) seems intended to ensure that the action in question is free.[7] Part (c) of Davidson's definition represents what we have called the agent's "better judgment," that is, the overall evaluation of his options contrary to which the incontinent agent acts.

Davidson notes that "there is no proving such actions exist; but it seems to me absolutely certain that they do" (p. 29). Why, then, is there is a persistent tendency, both in philosophy and in ordinary thought, to deny

that such actions are possible? Davidson's diagnosis is that two plausible principles which "derive their force from a very persuasive view of the nature of intentional action and practical reasoning" (p. 31) appear to entail that incontinence is impossible. He articulates those two principles as follows (p. 23):

> **P1**. If an agent wants to do *a* more than he wants to do *b* and he believes himself free to do either *a* or *b*, then he will intentionally do *a* if he does either *a* or *b* intentionally.
> **P2**. If an agent judges that it would be better to do *a* than to do *b*, then he wants to do *a* more than he wants to do *b*.

P2, Davidson observes, "connects judgements of what it is better to do with motivation or wanting" (p. 23); he adds later that it "states a mild form of internalism" (p. 26). Davidson is proposing, *contra* the extreme externalist position, that our evaluative judgments about the merits of the options we deem open to us are not motivationally inert. While he admits that one could quibble or tinker with the formulation of P1 and P2 (pp. 23–4; p. 27; p. 31), he is confident that they or something like them give expression to a powerfully attractive picture of practical reasoning and intentional action, one which assigns an important motivational role to the agent's evaluative judgments.[8]

The difficulty is, though, that P1 and P2—however attractive—together imply that an agent never intentionally does *b* when he judges that it would be better to do *a* (if he takes himself to be free to do either). And this certainly looks like a denial of the possibility of incontinent action. No wonder, then, that so many have been tempted to say that akratic action is impossible! Looking carefully, however, we can see that P1 and P2 do *not* imply the impossibility of incontinent actions *as Davidson has defined them*. For Davidson characterizes the agent who incontinently does *b* as holding, not that it would be better to do *a* than to do *b*, but that

it would be better, *all things considered*, to do *a* than to do *b*. Is the "all things considered" just a rhetorical flourish? Or does it mark a genuine difference between these two judgments? If these are two different judgments, and one can hold the latter without holding the former, then incontinent action is possible *even if P1 and P2 are true*.

In the rest of his paper Davidson sets out to vindicate that very possibility. The phrase "all things considered" is not, as it might seem, merely a minor difference in wording that allows weakness of will to get off on a technicality. Rather, that phrase marks an important contrast in logical form to which we would need to attend in any case in order properly to understand the structure of practical reasoning. For that phrase indicates a judgment that is *conditional* or *relational* rather than *all-out* or *unconditional* in form; and that difference is crucial.[9] We can better see the relational character of an all-things-considered judgment if we first look at evaluative judgments that play an important role in an earlier phase of practical reasoning, the phase where we consider what reasons or considerations favor doing *a* and what reasons or considerations favor doing *b*. (For simplicity I imagine a case in which an agent is choosing between only two mutually incompatible options, *a* and *b*.) These *prima facie* judgments, as Davidson terms them, take the form:

PF: In light of *r*, *a* is *prima facie* better than *b*.

In this schema *r* refers to a consideration, say that *a* would be relaxing, while *b* would be stressful. A PF judgment of this kind thus identifies one respect in which *a* is deemed superior to *b*, one perspective from which *a* comes out on top.

We should pause to note three things about PF judgments. (a) A PF judgment is not itself a conclusion in favor of the overall superiority of *a*. Such "all-out" evaluative judgments have a simpler logical form, namely:

AO: *a* is better than *b*.

(b) Indeed, no conclusion of the form AO follows logically from any PF judgment. (c) More strongly: the fact, taken by itself, that someone has made a certain PF judgment does not even supply her with sufficient grounds to draw the corresponding AO conclusion. For even if she makes one PF judgment which favors *a* over *b*, as in the case we imagined, she may *also* make *other* PF judgments which favor *b* over *a* (say, when *r* is the consideration that *b* would be lucrative, while *a* would be expensive). We do not want to say in that case that she has sufficient grounds to draw each of two incompatible conclusions (that *a* is better than *b*, and that *b* is better than *a*; these are incompatible provided the better-than relation is asymmetric, as I assume).

We have contrasted PF judgments with AO or "all-out" evaluative judgments. PF judgments are relational in character: they point out a *relation* which holds between the consideration *r* and doing *a*. (We could call that relation the "favoring" relation.) That relation is not such as to permit us to "detach" (as Davidson puts it, p. 37) an unconditional evaluative conclusion in favor of doing *a* from PF and the supposition that *r* obtains. That is, we are not to understand PF judgments as having the form of a material conditional.

Davidson's innovative suggestion is that judgments with this PF logical form are an appropriate way to model what happens in the early stages of practical reasoning, where we rehearse reasons for and against the options we are considering. And his stressing that no such PF judgment commits the agent to an overall evaluative conclusion in favor of *a* or *b* is useful in thinking about a case like Julie's ((1) above). We described Julie as knowing (and therefore believing) that *b* was more expensive than *a*, but opting for *b* nonetheless. We can imagine, then, that among the ingredients of Julie's practical reasoning was a PF judgment like this:

> In light of the fact that *b* is more expensive than *a*, *a* is *prima facie* better than *b*.

But this PF judgment alone, as we have seen, does not commit her to the overall judgment that *a* is better than *b*. For she may also have made other PF judgments, such as

> In light of the fact that *b* would be much more gastronomically exciting than *a*, *b* is *prima facie* better than *a*.

But we would not then want to say Julie has sufficient grounds to conclude that *a* is better than *b* and to conclude that *b* is better than *a*. She does not have sufficient grounds to embrace a contradiction; her premises all seem consistent. So her various PF judgments, when considered separately, must *not* each commit her to a corresponding overall conclusion in favor of *a* or *b*.

Practical reasoning, Davidson suggests, starts from judgments like these, each identifying one respect in which one of the options is superior. But in order to make progress in our practical reasoning we shall eventually need to consider how *a* compares to *b* not just with respect to *one* consideration, but in the light of several considerations taken together. That is, Julie will eventually need to consider how to fill in the blanks in a PF judgment like this:

> In light of the fact that *b* is more expensive than *a and* the fact that *b* would be much more gastronomically exciting than *a*, … is *prima facie* better than ….

This PF judgment is more *comprehensive* than the ones we attributed to Julie a moment ago, as it takes into account a broader range of considerations. (I take the label "comprehensive" from Lazar 1999.) Now in Julie's case we can surmise how she filled in those blanks: with "*b* is

prima facie better than *a*." Julie's filling in the blanks in that way can naturally be taken as expressing the view that the much greater gastronomic excitement promised by *b outweighs* or *overrides b*'s inferiority to *a* from a strictly financial standpoint.

We can generalize our schema for PF judgments to account for the possibility of relativizing our comparative assessment of *a* and *b* not just to a single consideration, but to multiple considerations taken together or as a body:

PFN: In light of $\langle r_1, \ldots, r_n \rangle$, *a* is *prima facie* better than *b*.

Notice that PFN judgments are still relational in form: they assert that a relation (the "favoring" relation) holds between the *set* of considerations $\langle r_1, \ldots, r_n \rangle$ and doing *a*. Indeed, the relational character of a PFN judgment remains even if we make it as comprehensive as we can: if we expand the set $\langle r_1, \ldots, r_n \rangle$ to incorporate *all* the considerations the agent deems relevant to her decision. Following Davidson (p. 38), let us give the label *e* to that set. So even the following judgment:

ATC: In light of *e*, *a* is *prima facie* better than *b*.

is a relational or conditional judgment and not an all-out conclusion in favor of doing *a*. To make a judgment of the form ATC is *not* to draw an overall conclusion in favor of doing *a*.

We may be better able to see this by considering an analogy from theoretical reason. Suppose Hercule Poirot has been called in to investigate a murder. We can imagine him assessing bits of evidence as he encounters them:

> In light of the fact that the murder weapon belongs to Colonel Mustard, Mustard looks guilty;

> In light of his having an alibi for the time of the murder, Mustard looks not guilty;

and so on. These are theoretical analogues of the PF judgments relativized to single considerations which we looked at earlier. However, Poirot will eventually need to consider how these various bits of evidence add up; that is, he will eventually need to fill in the blanks in a more comprehensive PFN judgment like this:

> In light of $\langle e_1, \ldots, e_n \rangle$, … looks to be the guilty party,

where $\langle e_1, \ldots, e_n \rangle$ is a set of bits of pertinent evidence. Notice, though, that no such PFN judgment actually constitutes settling on a particular person as the culprit. For even if we put in a maximally large $\langle e_1, \ldots, e_n \rangle$ consisting of *all* the evidence Poirot has seen, and imagine him thinking

> All the evidence I have seen points toward Colonel Mustard as the guilty party,

to make this observation is manifestly *not* to conclude that Mustard is guilty.

In the same way, an ATC or all-things-considered judgment, although comprehensive, is still relational in nature, and therefore distinct from an AO judgment in favor of a. That is, it is possible to make an ATC judgment in favor of a without making the corresponding AO judgment in favor of a. (This is the analogue of Poirot's position.) And this is the key to Davidson's solution to the problem of how weakness of will is possible. For ATC is, precisely, *the agent's better judgment* as Davidson construes it in his definition of incontinent action. P1 and P2 together imply that an agent who reaches an AO conclusion in favor of a will not intentionally do b. But the incontinent agent never reaches such an AO conclusion. With respect to a, he remains stuck at the Hercule Poirot

stage: he sees that the considerations he has rehearsed, taken as a body, favor *a*, but he is unwilling or unable to make a commitment to *a* as the thing to do.[10] He makes only a relational ATC judgment in favor of *a*, contrary to which he then acts.

What should we say about an agent who does this? Returning to the three features of *prima facie* or PF judgments which we noted earlier, features (a) and (b) hold even of the special subclass of PF judgments which are ATC judgments. Such judgments neither are equivalent to, nor logically imply, any AO judgment. So the incontinent agent who fails to draw the AO conclusion which corresponds to his ATC conclusion, and to perform the corresponding action, is not committing "a simple logical blunder" (p. 40). Notably, he does not contradict himself. He does, however, exhibit a defect in rationality, on Davidson's account. For feature (c) of PF judgments in general does *not* hold of the special subclass of such judgments which are ATC judgments. Drawing an *ATC* conclusion in favor of *a does* give one sufficient grounds to conclude that *a* is better *sans phrase* and, indeed, to do *a*. For Davidson proposes that the transition from an ATC judgment in favor of *a* to the corresponding AO judgment, and to doing *a*, is enjoined by a substantive principle of rationality which he dubs "the principle of continence." That principle tells us to "perform the action judged best on the basis of all available relevant reasons" (p. 41); and the incontinent agent violates this injunction. The principle of continence thus substantiates the idea that "what is wrong is that the incontinent man acts, and judges, irrationally, for this is surely what we must say of a man who goes against his own best judgement" (p. 41). He acts irrationally in virtue of violating this substantive principle, obedience to which is a necessary condition for rationality.

We must put this point about the irrationality of incontinence with some care, however. For recall that an incontinent action must itself be

intentional, that is, done for a reason. The weak-willed agent, then, has a reason for doing *b*, and does *b* for that reason. What he lacks—and lacks by his own lights—is a *sufficient* reason to do *b*, given all the considerations that he takes to favor *a*. As Davidson puts it, if we ask "what is the agent's reason for doing [*b*] when he believes it would be better, all things considered, to do another thing, then the answer must be: for this, the agent has no reason" (p. 42). And this is so even though he does have a reason for doing *b* (p. 42, n. 25). Because the agent has, by his own lights, no adequate reason for doing *b*, he cannot make sense of his own action: "he recognizes, in his own intentional behaviour, something essentially surd" (p. 42). So akratic action, while *possible* on Davidson's account, is nonetheless necessarily *irrational*; this is the sense in which it is a defective and not fully intelligible instance of agency, despite being a very real phenomenon.

3. The Debate After Davidson

3.1 Internalist and Externalist Strands

Davidson has certainly presented an arresting theory of practical reasoning. But has he shown how weakness of the will is possible? Most philosophers writing after him, while acknowledging his pathbreaking work on the issue, think he has not. One principal difficulty which subsequent theorists have seized on is that Davidson's view can account for the possibility of action contrary to one's better judgment *only if* one's better judgment is construed merely as a conditional or *prima facie* judgment. Davidson's P1 and P2 in fact rule out the possibility of free intentional action contrary to an all-out or unconditional evaluative judgment.[11] But it seems that such cases exist. Michael Bratman, for instance, introduces us to Sam, who, in a depressed state, is deep into a bottle of wine, despite his acknowledged need for an early wake-up and a clear head tomorrow. Sam's friend, stopping by, says:

> "Look here. Your reasons for abstaining seem clearly stronger than your reasons for drinking. So how can you have thought that it would be best to drink?" To which Sam replies: "I don't think it would be best to drink. Do you think I'm stupid enough to think that, given how strong my reasons for abstaining are? I think it would be best to abstain. Still, I'm drinking." (1979, p. 156)

Sam's case certainly seems possible as described. Davidson's view, though, must reject it as impossible. Given his conduct, Sam can't think it best to abstain; at most, he thinks it all-things-considered best to abstain, a very different kettle of fish. But this seems false of Sam: there is no evidence that he has remained stuck at the Hercule Poirot stage with respect to the superiority of abstaining. He seems to have gone all the way to a judgment *sans phrase* that abstaining would be better; and yet he drinks.

Ironically, this complaint makes Davidson out to be a bit like Hare. Like Hare, Davidson subscribes to an internalist principle (P2) which connects evaluative judgments with motivation and hence with action. (Indeed, in light of the difficulty raised here, one might wonder if Davidson is entitled to consider P2 a "mild" form of internalism (p. 26).) As with Hare, this internalist commitment rules out as impossible certain kinds of action contrary to one's evaluative judgment. Now Davidson, like Hare, does accept the possibility of certain phenomena in this neighborhood; but—as with Hare—critics think the cases permitted by his analysis simply do not exhaust the range of actual cases of weakness of will. The phenomenon seems to run one step ahead of our attempts to make room for it.

Those writing after Davidson have tended to focus, then, on the question of the possibility and rational status of action contrary to one's *unconditional* better judgment.[12] With respect to these questions, it

seems to me, the challenge sketched at the end of Section 1 above remains in full force. What is required is a view which successfully navigates between the Scylla of an extreme internalism about evaluative judgment which would preclude the possibility of weakness of will, and the Charybdis of an extreme externalism which would deny any privileged role to evaluative judgment in practical reasoning or rational action. For one's verdict about *akrasia* will in general be closely connected to one's more general views of action, practical reasoning, rationality, and evaluative judgment—as was certainly true of Davidson.

Naturally, different theorists have plotted different courses through these shoals. Some tack more to the internalist side, wishing to preserve a strong internal connection between evaluation and action even at the risk of denying or seeming to deny the possibility of akratic action (or at least some understandings of it). Examples of some post-Davidson treatments which share a broadly internalist emphasis, even if they feature different flavors of internalism, are those of Bratman (1979), Buss (1997), Tenenbaum (1999), and Stroud (2003). The main danger for such approaches is that in seeking to preserve and defend a certain picture of the primordial role of evaluative thought in rational action—a picture critics are likely to dismiss as too rationalistic—such theorists may be led to reject common phenomena which ought properly to have constrained their more abstract theories. (See the opening of Wiggins 1979 for a forceful articulation of this criticism.)

Other theorists, by contrast, are more drawn toward the externalist shoreline. They emphasize the motivational importance of factors other than the agent's evaluative judgment and the divergences that can result between an agent's evaluation of her options and her motivation to act. They are thus disinclined to posit any strong, necessary link between evaluative judgment and action. Michael Stocker, for instance, argues that the philosophical tradition has been led astray in assuming that evaluation

dictates motivation. "Motivation and evaluation do not stand in a simple and direct relation to each other, as so often supposed," he writes. Rather, "their interrelations are mediated by large arrays of complex psychic structures, such as mood, energy, and interest" (1979, pp. 738–9). Similarly, Alfred Mele proposes as a fundamental and general truth—and one that underlies the possibility of *akrasia*—that "the motivational force of a want may be out of line with the agent's evaluation of the object of that want" (1987, p. 37). Mele goes on to offer several different reasons why the two can come apart: for example, rewards perceived as *proximate* can exert a motivational influence disproportionate to the value the agent reflectively attaches to them (1987, ch. 6). Such wants may function as strong *causes* even if the agent takes them to constitute weak *reasons*.

Views that downplay the role of evaluative judgment in action and hence tack more toward the externalist side of the channel may more easily be able to accept the possibility and indeed the actuality of weakness of will. But they are subject to their own challenges. For example, suppose we follow Mele's image of *akrasia* and posit that a certain agent is caused to do x by motivation to do x which is dramatically out of kilter with her assessment of the merits of doing x. In what sense, then, is her doing x free, intentional, and uncompelled? Such an agent might seem rather to be at the mercy of a motivational force which is, from her point of view, utterly alien. Thus, worries about distinguishing *akrasia* from compulsion come back in full force in connection with proposals like these. (See fn. 7 above for relevant references; Buss and Tenenbaum press these worries against accounts like Mele's in particular.) Moreover, there is the danger, for accounts of this more externalist stripe, of taking too much of the mystery out of weakness of will. Even if akratic action is possible and indeed actual, it remains a puzzling, marginal, somehow defective instance of agency, one that we rightly find not fully intelligible. Views that do not assign a privileged place in rational deliberation and action to

the agent's overall assessment of her options risk making akratic action seem no more problematic than Julie's or Jimmy's decisions, or Hare's agent who fails to pick up the roundest stone in the vicinity.

3.2 Weakness of Will as Potentially Rational

The "externalist turn" toward downplaying the role of an agent's better judgment and emphasizing other psychic factors instead is connected to a second way in which some theorists writing after Davidson have dissented from his analysis. Davidson, as we saw, viewed akratic action as possible, but irrational. The weak-willed agent acts contrary to what she herself takes to be the balance of reasons; her choice is thus unreasonable by her own lights. On this picture, incontinent action is a paradigm case of practical irrationality. Many other theorists have agreed with Davidson on this score and have taken *akrasia* to be perhaps the clearest example of practical irrationality. But some writers (notably Audi 1990, McIntyre 1990, and Arpaly 2000) have questioned whether akratic action *is* necessarily irrational. Perhaps we ought to leave room, not just for the *possibility* of akratic action, but for the potential *rationality* of akratic action.

The irrationality which is held necessarily to attach to akratic action derives from the discrepancy between what the agent judges to be the best (or better) thing to do, and what she does. That is, her action is faulted as irrational in virtue of not conforming to her better judgment. But—ask these critics—what if her better judgment is itself faulty? There is nothing magical about an agent's better judgment that ensures that it is correct, or even warranted; like any other judgment, it can be in error, or even unjustified. (Recall that by "better judgment" we meant, all along, only "a judgment as to which course of action is better," not "a *superior* judgment.") Where the agent's better judgment is itself defective, in doing what she deems herself to have insufficient reason to do, the agent may

actually be doing what she has most reason to do. "Even though the akratic agent does not believe that she is doing what she has most reason to do, it may nevertheless be the case that the course of action that she is pursuing is the one that she has … most reason to pursue" (McIntyre 1990, p. 385). In that sense the akratic agent may be wiser than her own better judgment.

How, concretely, could an agent's better judgment go astray in this way? Perhaps her survey of what she took to be the relevant considerations did not include, or did not attach sufficient weight to, what were in fact significant reasons in favor of one of the possible courses of action. She may have overlooked these, or (wrongly) deemed them not to be reasons, or failed to appreciate their full force; and in that case her judgment of what it is best to do will be incorrect. Consider, for example, Jonathan Bennett's Huckleberry Finn (Bennett 1974, discussed in McIntyre 1990), who akratically fails to turn in his slave friend Jim to the authorities. Huck's judgment that he ought to do so, however, was based primarily on what he took to be the force of Miss Watson's property rights; it ignored his powerful feelings of friendship and affection for Jim, as well as other highly relevant factors. His "better judgment" was thus not in fact a very comprehensive judgment; it did not take into account the full range of relevant considerations.

Or consider Emily, who has always thought it best that she pursue a Ph.D. in chemistry (Arpaly 2000, p. 504). When she revisits the issue, as she does periodically, she discounts her increasing feelings of restlessness, sadness, and lack of motivation as she proceeds in the program, and concludes that she ought to persevere. But in fact she has very good reasons to quit the program—her talents are not well suited to a career in chemistry, and the people who are thriving in the program are very different from her. If she impulsively, akratically quits the program, purely on the basis of her feelings, Emily is in fact doing just what she

ought to do.[13] That her action conflicts with her better judgment does not significantly impugn its rationality, given all the considerations that *do* support her quitting the program. "A theory of rationality should not assume that there is something special about an agent's best judgment. An agent's best judgment is just another belief" (Arpaly 2000, p. 512). Emily's action conflicts, then, with one belief she has; but it coheres with many more of her beliefs and desires overall. So even though she may find her own action inexplicable or "surd," she is in fact acting rationally, although she does not know it. *Contra* Davidson, "we can … act rationally just when we cannot make any sense of our actions" (Arpaly 2000, p. 513).

It is unclear, however, whether these arguments and examples are likely to sway those who take *akrasia* to be a paradigm of practical irrationality. These dissenters stress the *substantive merits* of the course of action the akratic agent follows. But traditionalists may say that is beside the point: however well things turn out, the practical thinking of the akratic agent still exhibits a *procedural defect*. Someone who flouts her own conclusion about where the balance of reasons lies is *ipso facto* not reasoning well. Even if the action she performs is in fact supported by the balance of reasons, she does not think it is, and that is enough to show her practical reasoning to be faulty. The defenders of the traditional conception of *akrasia* as irrational thus wish to grant special rational authority (in this procedural sense) to the agent's better judgment, even if they admit that such a judgment can be substantively incorrect. By contrast, the dissenters "[do] not believe best judgments have any privileged role" (Arpaly 2000, p. 513). We see again the contrast between "internalist" and "externalist" tendencies in the debates over weakness of will.

3.3 Changing the Subject

A final revisionist strand now emerging in the literature takes the agent's

better judgment even farther out of the picture. In an outstandingly lucid and stimulating essay published in 1999, Richard Holton argued that weakness of will is not action contrary to one's better judgment at all. The literature has gone astray in understanding weakness of will in this way; weakness of will is actually quite a different phenomenon, in which the agent's better judgment plays no role.[14] For Holton, when ordinary people speak of weakness of will they have in mind a certain kind of failure to act on one's *intentions*. What matters for weakness of will, then, is not whether you deem another course of action superior at the time of action. It is whether you are abandoning an intention you previously formed. Weakness of will as the untutored understand it is not *akrasia* (if we reserve that term for action contrary to one's better judgment), but rather a certain kind of failure to stick to one's plans. This understanding of weakness of will changes the subject in two ways. First, the state of the agent with which the weak-willed action is in conflict is not an evaluative judgment (as in *akrasia*) but a different kind of state, namely an intention. Second, it is not essential that there be *synchronic* conflict, as *akrasia* demands. You must act contrary to your *present* better judgment in order to exhibit *akrasia*; conflict with a *previous* better judgment does not indicate *akrasia*, but merely a change of mind. However, you can exhibit weakness of will as Holton understands it simply by abandoning a previously formed intention.

Of course not all cases of abandoning or failing to act on a previously formed intention count as weakness of will. I intend to run five miles tomorrow evening. If I break my leg tomorrow morning and fail to run five miles tomorrow evening, I will not have exhibited weakness of will. How can we characterize *which* failures to act on a previously formed intention count as weakness of will? Holton's answer has two parts. First, he says, there is an irreducible normative dimension to the question whether someone's abandoning of an intention constituted weakness of will (Holton 1999, p. 259). That is, there is no purely descriptive criterion

(such as whether her action conflicted with her better judgment) which is sufficient for weakness of will; in order to decide whether a given case was an instance of weakness of will we must consider normative questions, such as whether it was *reasonable* for the agent to have abandoned or revised that intention, or whether she *should* have done so. In the case of my broken leg, for instance, it was clearly reasonable for me to abandon my intention; that is why I could not be charged with weakness of will in that case.

Second, says Holton, we need to attend to an important subclass of our intentions to do something at a future time, namely *contrary-inclination-defeating intentions*, or, as he later terms them (Holton 2003), *resolutions*. Resolutions are intentions that are formed precisely in order to insulate one against contrary inclinations one expects to feel when the time comes. Thus one reason I might form an intention on Monday to run five miles on Tuesday—as opposed to leaving the issue open until Tuesday, for decision then—is to reduce the effect of feelings of lassitude to which I fear I may be subject when Tuesday rolls around. Then suppose Tuesday rolls around; I am indeed prey to feelings of lassitude; and I decide as a result not to run. *Now* I can be charged with weakness of will. Weakness of will involves, specifically, a failure to act on a *resolution*; this is sufficient to differentiate weakness of will from mere change of mind and even from caprice (which is a *different* species of unreasonable intention revision, according to Holton).

As a recent paper by Alison McIntyre shows (McIntyre 2006), understanding weakness of will in this way casts a fresh light on the issue of its rational status. The weak-willed agent abandons a resolution because of a contrary inclination of exactly the type which the resolution was expressly designed to defeat. Therefore, as McIntyre underlines, weak-willed action always involves a procedural rational defect:[15] a technique of self-management has been deployed but has failed

(McIntyre 2006, p. 296). To that extent we have grounds to criticize weak-willed action simply in virtue of the second of the ways in which Holton wishes to distinguish weakness of will from a mere change of mind, without even resolving the potentially murky issue of whether the agent was *reasonable* in abandoning her intention.

McIntyre holds, however, that it would be overstating the case to say that because weakness of will involves this procedural defect, it is always irrational (McIntyre 2006, p. 290; pp. 298–9; p. 302). She proposes rather that practical rationality has multiple facets and aims, and that failure in one respect or along one dimension does not automatically justify the especially severe form of rational criticism which we intend by the term "irrational." For example, consider an agent who succumbs to contrary inclination of exactly the type expected when the time comes to act on a truly *stupid* resolution. (Holton gives the example of resolving to go without water for two days just to see what it feels like: Holton 2003, p. 42.) There will indeed be a blemish on this agent's rational scorecard if he eventually gives in and drinks: he will have failed in his attempt at self-management. But wouldn't it be rationally far *worse* for him to stick to his silly resolution no matter what the cost?

We can also re-examine the issue of the rationality of *akrasia* in light of this analysis of weakness of will; for we can distinguish between akratic and non-akratic cases of the latter. As McIntyre points out, resolutions typically rest on judgments about what it is best that one do at a (future) time t. If an agent fails to act on a previously formed resolution to do a at t, thus exhibiting weakness of will, we can distinguish the case in which he still endorses at t the judgment that it is best that he do a at t (even though he does not do it) from the case in which he abandons that judgment as well as his resolution. In the latter, non-akratic type of case, the agent in effect rationalizes his failure to live up to his resolution by deciding that it is not after all best that he do a at t. McIntyre points out

that the traditional view that *akrasia* is always irrational seems to give us a perverse incentive to rationalize, since in that case we escape the grave charge of practical irrationality, being left only with the procedural practical defect present in all cases of weakness of will (McIntyre 2006, p. 291). But this seems implausible: are the two sub-cases so radically different in their rational status? Indeed, she argues, if anything, akratic weakness of will is typically rationally *preferable* to rationalizing weakness of will (McIntyre 2006, p. 287; pp. 309ff.). "In the presence of powerful contrary inclinations that bring about a failure to be resolute," she writes, "resisting rationalization and remaining clearheaded about one's reasons to act can constitute a modest accomplishment" (McIntyre 2006, p. 311). Have we witnessed the transformation of *akrasia* from impossible, to irrational, to downright admirable?

Bibliography

- Aristotle, *Nicomachean Ethics*, book VII, chs. 1-10.
- Arpaly, N., 2000, "On Acting Rationally Against One's Better Judgment," *Ethics* **110**: 488-513.
- Audi, R., 1979, "Weakness of Will and Practical Judgment," *Noûs* **13**: 173-196.
- ——, 1990, "Weakness of Will and Rational Action," *Australasian Journal of Philosophy* **68**: 270-281.
- Austin, J. L., 1956/7, "A Plea for Excuses," in Austin 1979, pp. 175-204.
- ——, 1979, *Philosophical Papers*, 3rd ed., J. O. Urmson and G. J. Warnock (eds.), Oxford: Oxford University Press.
- Bennett, J., 1974, "The Conscience of Huckleberry Finn," *Philosophy* **49**: 123-134.
- Bobonich, C., and Destrée, P. (eds.), 2007, Akrasia *in Greek Philosophy: From Socrates to Plotinus*, Leiden, Boston: Brill.
- Bratman, M., 1979, "Practical Reasoning and Weakness of the Will,"

Noûs **13**: 153-171.

- Buss, S., 1997, "Weakness of Will," *Pacific Philosophical Quarterly* **78**: 13-44.
- Charlton, W., 1988, *Weakness of Will*, Oxford: Basil Blackwell.
- Davidson, D., 1970, "How Is Weakness of the Will Possible?," in Davidson 1980, pp. 21-42.
- ——, 1978, "Intending," in Davidson 1980, pp. 83-102.
- ——, 1980, *Essays on Actions and Events*, Oxford: Clarendon Press.
- ——, 1982, "Paradoxes of Irrationality," in Davidson 2004, pp. 169-187.
- ——, 2004, *Problems of Rationality*, Oxford: Clarendon Press.
- Dunn, R., 1987, *The Possibility of Weakness of Will*, Indianapolis: Hackett.
- Gosling, J., 1990, *Weakness of the Will*, London and New York: Routledge.
- Hare, R. M., 1952, *The Language of Morals*, Oxford: Clarendon Press.
- ——, 1963, *Freedom and Reason*, Oxford: Clarendon Press.
- ——, 2001, "Weakness of Will," in *The Encyclopedia of Ethics*, 2nd ed., L. Becker and C. Becker (eds.), New York: Routledge, pp. 1789-1792.
- Hill, T., 1986, "Weakness of Will and Character," in Hill 1991, pp. 118-137.
- ——, 1991, *Autonomy and Self-Respect*, Cambridge: Cambridge University Press.
- Hoffmann, T. (ed.), 2008, *Weakness of Will from Plato to the Present*, Washington: Catholic University of America Press.
- Holton, R., 1999, "Intention and Weakness of Will," *Journal of Philosophy* **96**: 241-262.
- ——, 2003, "How is Strength of Will Possible?," in Stroud and Tappolet 2003, pp. 39-67.

- Jackson, F., 1984, "Weakness of Will," *Mind* **93**: 1-18.
- Jones, K., 2003, "Emotion, Weakness of Will, and the Normative Conception of Agency," in A. Hatzimoysis (ed.), *Philosophy and the Emotions*, Cambridge: Cambridge University Press, pp. 181-200.
- Kennett, J., 2001, *Agency and Responsibility: A Common-Sense Moral Psychology*, Oxford: Clarendon Press.
- Lazar, A., 1999, "Akrasia and the Principle of Continence or What the Tortoise Would Say to Achilles," in L. E. Hahn (ed.), *The Philosophy of Donald Davidson* (Library of Living Philosophers, **27**), Chicago: Open Court, pp. 381-401.
- McIntyre, A., 1990, "Is Akratic Action Always Irrational?," in *Identity, Character, and Morality*, O. Flanagan and A. Rorty (eds.), Cambridge, MA: MIT Press, pp. 379-400.
- ——, 2006, "What Is Wrong With Weakness of Will?," *Journal of Philosophy* **103**: 284-311.
- Mele, A., 1987, *Irrationality*, New York: Oxford University Press.
- ——, 1991, "Akratic Action and the Practical Role of Better Judgment," *Pacific Philosophical Quarterly* **72**: 33-47.
- ——, 2002, "Akratics and Addicts," *American Philosophical Quarterly* **39**: 153-167.
- Pears, D., 1984, *Motivated Irrationality*, Oxford: Clarendon Press.
- Plato, *Protagoras*, in *The Collected Dialogues of Plato*, E. Hamilton and H. Cairns (eds.), Princeton: Princeton University Press, 1961, pp. 308-352.
- Rorty, A., 1980, "Where Does the Akratic Break Take Place?," *Australasian Journal of Philosophy* **58**: 333-347.
- Saarinen, R., 1994, *Weakness of the Will in Medieval Thought: From Augustine to Buridan*, Leiden, New York: Brill.
- Smith, M., 2003, "Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion," in Stroud and Tappolet 2003, pp. 17-38.

- Stocker, M., 1979, "Desiring the Bad: An Essay in Moral Psychology," *Journal of Philosophy* **76**: 738-753.
- Stroud, S., 2003, "Weakness of Will and Practical Judgement," in Stroud and Tappolet 2003, pp. 121-146.
- Stroud, S., and Tappolet, C. (eds.), 2003, *Weakness of Will and Practical Irrationality*, Oxford: Clarendon Press.
- Tappolet, C., 2003, "Emotions and the Intelligibility of Akratic Action," in Stroud and Tappolet 2003, pp. 97-120.
- Tenenbaum, S., 1999, "The Judgment of a Weak Will," *Philosophy and Phenomenological Research* **59**: 875-911.
- Thero, D., 2006, *Understanding Moral Weakness*, Amsterdam, New York: Rodopi.
- Walker, A., 1989, "The Problem of Weakness of Will," *Noûs* **23**: 653-676.
- Wallace, R. J., 1999, "Addiction as Defect of the Will: Some Philosophical Reflections," *Law and Philosophy* **18**: 621-654.
- Watson, G., 1977, "Skepticism About Weakness of Will," *Philosophical Review* **86**: 316-339.
- Wiggins, D., 1979, "Weakness of Will, Commensurability, and the Objects of Deliberation and Desire," *Proceedings of the Aristotelian Society* **79**: 251-277.
- Wilkerson, T. E., 1997, *Irrational Action: A Philosophical Analysis*, Aldershot: Ashgate.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

action | Aristotle, General Topics: ethics | Davidson, Donald | practical reason | practical reason: and the structure of actions | practical reason:

medieval theories of

Acknowledgments

I am grateful to Eric Guindon for very useful research assistance.

Notes to Weakness of Will

1. Until section 3.3 of this essay, where I discuss some revisionist doubts about understanding weakness of will as action contrary to one's better judgment, I will follow the tradition in using the expression "weakness of will" to refer to the phenomenon I have just described. Note that while the phrase "weakness of will" might suggest a general character trait, contemporary philosophers have generally concentrated on weak-willed *action*, that is, on individual actions performed against the agent's better judgment. (See Hill 1986 for discussion of weakness of will as a character trait.) They have often used "incontinent" and "akratic" as synonyms for "weak-willed" in this sense, and until section 3.3 I shall follow tradition in these respects too. ("Akratic" comes from the Greek *akrasia*, lack of self-control.)

2. The present essay focuses on the contemporary (post-World War II) literature on weakness of will; the reader interested in tracing the history of philosophical discussion of this topic is invited to consult the following additional sources. Within the *Stanford Encyclopedia of Philosophy*, the entry on *Aristotle, General Topics, ethics* discusses his views on *akrasia* at some length, and the entry on *practical reason, medieval theories of* provides pertinent background on medieval views of practical reasoning and the will. Among monographs on weakness of will, those by Charlton (1988) and Gosling (1990) have a significant historical component: Charlton's chs. 2 and 3, and Gosling's chs. I–VII, discuss ancient (and, in the case of Gosling, medieval and early modern) work on the topic prior

to engaging with the contemporary literature in later chapters. See also, in similar guise, Wilkerson 1997, ch. 1, and Thero 2006, chs. 2–5. Among works more purely historical in nature, Saarinen 1994 is a monograph devoted to medieval treatments of the problem. Bobonich and Destrée 2007 is a new collection of essays on ancient discussions of *akrasia*, the introduction to which gives a useful overview. Another new collection, Hoffmann 2008, discusses various treatments of weakness of will "from Plato to the present," with detailed examination of a number of medieval authors.

3. Hare sometimes uses weaker terms than "entail" in this connection: see Hare 1952, p. 172 and Hare 1963, p. 56. But his subsequent conclusions follow only if he did truly mean that an action-guiding judgment must *entail* an answer to the practical question.

4. Hare makes more subtle distinctions among a range of possible cases in Hare 1963, ch. 5, and Hare 2001, but these are the two principal categories.

5. Davidson 1970, p. 22. For consistency with the rest of the text I have re-lettered Davidson's variables as *a* and *b* in this quotation and in principles P1 and P2, quoted below.

6. If I decide it would be best not to turn on the light when checking on my sleeping baby, but then a violent sneeze causes my arm to fly up and hit the light switch, thus turning on the light, this ought not to count as an incontinent action on my part; and it does not, on Davidson's criteria. Furthermore, condition (a) rules out cases in which, say, a man thinks it best that he not send a valentine to Margery Morningstar, but intentionally sends a valentine to Margery Eveningstar, not knowing that Margery Eveningstar *is* Margery Morningstar (Davidson 1970, p. 25). Such an act should not count as an instance of *akrasia* either; and it does not, on Davidson's criteria, since it is not intentional under the description

"sending a valentine to Margery Morningstar."

7. Davidson seems to endorse the idea that incontinent action must be free (p. 22, n. 1) and, in particular, uncompelled (p. 29), even if his official characterization does not use these terms explicitly. Whether and how weakness of will can be distinguished from compulsion has been a subject of much debate in the literature: see Gary Watson's classic paper on the subject (1977) and also Audi 1979, Mele 1987, ch. 2, Buss 1997, Tenenbaum 1999, Wallace 1999, Kennett 2001, ch. 6, Mele 2002, and Smith 2003 for discussion of this issue.

8. Davidson further sketches that picture in section II of Davidson 1970 and in Davidson 1978.

9. Davidson explains the logical distinction between these two types of judgment at 1970, pp. 37–41; my presentation is also indebted to Bratman 1979. Davidson typically calls the first type of judgment "*prima facie* judgments," a usage I shall follow.

10. Indeed, not only does he fail to draw an AO conclusion in favor of *a*; he actually draws an AO conclusion in favor of *b*. This implication of Davidson's treatment emerges with greater clarity from Davidson 1978, in which he identifies intentions with all-out or unconditional evaluative judgments. (Recall that the incontinent agent does *b intentionally*.)

11. Because Davidson does not, strictly speaking, commit himself to P1 and P2 in the precise formulations given in his 1970 (see pp. 23–4; p. 27; p. 31), it is overstating the case to say that his analysis *rules out* action contrary to an unconditional judgment. At a minimum, though, such action (if possible) is wholly unaccounted for by his treatment, and in light of the totality of the evidence in his 1970 and 1978 it seems reasonable to conclude that he does view it as impossible. I will heretofore speak as if Davidson really had committed himself to the truth

of P1 and P2, and hence to the impossibility of such action.

12. Alfred Mele, for instance, focuses on what he calls "strict akratic action" in his 1987 (see esp. chs. 1–3 and 6), and David Pears explores "last-ditch *akrasia*" at length in his 1984 (see esp. chs. VI–X).

13. Other writers who have especially stressed the role of emotions in akratic action and indeed in *rational* akratic action include Jones 2003 and Tappolet 2003. In doing so they reprise an idea mooted by the later Davidson (1982), namely that we can make the most sense of *akrasia* when it involves distinct sub-systems within the mind operating to some degree autonomously. (One of the distinctive features of emotions as psychic forces, according to such writers, is that they arise and influence us relatively independently of conscious judgment.)

14. See Jackson 1984 for a different revisionist understanding of weakness of will which also bypasses the agent's better judgment.

15. McIntyre implicitly takes her earlier self to task for having neglected the procedural aspect of rationality and equated what is rational with what we have most reason to do (McIntyre 2006, p. 289; p. 299).