Forthcoming in T. O'Connor and C. Sandis, *A Companion to the Philosophy of Action* (Blackwell).

AKRASIA AND IRRATIONALITY

Sergio Tenenbaum

University of Toronto

Akrasia and *accidie* are traditionally recognized as tow of the clearest cases of practical irrationality. An "akratic action", to a first approximation, is an intentional action that the agent recognizes to be in conflict with what she judges to be the best course of action. So an agent who continues smoking even though she thinks it would be better if she were to quit smoking would be engaging in akratic actions. In a state of *accidie*, by contrast, the agent recognizes that there is something of value that he can and ought to bring about, and yet he does not engage in any action to bring it about, or in any other course of action that he judges he ought to undertake, or in any other course of action that he judges to be more or equally valuable. So, for instance, a depressed agent who knows that he can go to work and help support his family but stays in bed nonetheless is suffering from *accidie*. *Akrasia*, *accidie*, and other forms of practical irrationality are philosophically interesting in themselves, but they are also phenomena that are taken to be test cases for various philosophical theories in the realm of ethics and practical reason. For instance, the acceptance of the "guise of the good" thesis (the view that all intentional action aims at the good) is often taken to be incompatible with the possibility of *akrasia* or *accidie* (see (Stocker 1979); for an argument that the guise of the good thesis is

compatible with the possibility of *akrasia* and *accidie*, see (Tenenbaum 2007)); ethical internalism, the view that moral judgments necessarily motivate, is often taken to be incompatible with the possibility of *accidie* (Smith 1994). More generally, some philosophers have argued that theories of rationality, such as the view that rationality only commands that one takes the means to one's given ends, leave no room for the possibility of practical irrationality (Korsgaard 1997, Pears 1982).

These kinds of claims are often put forward on the assumption that it is a matter of empirical fact that phenomena such as *akrasia* and *accidie* exist, but the grounds for this assumption are not clear (Korsgaard is a notable exception, since she thinks that the possibility of irrationality is constitutive of the norms of rationality). An action can count as a case of *akrasia* or *accidie*, if the action (or inaction) that the agent chooses conflicts with some kind of evaluative judgment the agent makes. However, it is not immediately obvious why we must attribute the relevant evaluative judgment to the agent in question. Philosophers often rely on the supposition that agents in these situations would sincerely assent to certain agents. But agents might be self-deceived, confused, or simply changing their minds. Given the yeomen's work that the possibility of such irrational behaviour is supposed to perform, one would expect that philosophers would have done a better job of showing that "akrasia" and "accidie" denote real phenomena.

References to weakness of will are common in ordinary parlance, but doubtless the words "akrasia" and "accidie" are terms of art (at least in English). Since philosophers have traditionally assumed that the ordinary notion of weakness

of will and the philosophical conception of *akrasia* coincide, they could at least draw some comfort from the fact that it is a well-entrenched part of our ordinary understanding of agents that they can be irrational exactly by exhibiting weakness of will. However, it has been recently suggested that what is ordinarily described as weakness of will is *not* the same phenomenon as the one covered by the philosopher's notion of *akrasia*. In what follows, I'll try first to argue that, insofar as the ordinary notion of weakness of will denotes some kind of irrationality in the agent, the traditional view that identifies weakness of will with *akrasia* is the correct one. I then try to suggest more general and more promising ways of establishing the reality of such phenomena. I'll focus on *akrasia*, but much of what I say, especially in the second part of the argument, should apply to *accidie* too.

Davidson's classic paper on weakness of will (Davidson 1980) defines weakness of will as failure to act in accordance to what one acknowledges (or at least thinks) to be the correct evaluative judgment. That is, a weak-willed agent is one who judges that A is better than B all-things-considered, but (freely) chooses B over A. Versions of this view seem to have been endorsed by historical figures from Aristotle to Kant (((Aristotle 1985), (Kant, et al. 1998)). On Davidson's explication of *akrasia*, the agent's privileged evaluative judgment is identified with what he calls an "all-things-considered judgment". The all-things-considered judgment is best understood when contrasted with *prima-facie* evaluative judgments such as "insofar as A, but not B, will cause me to feel pleasure, A is better than B". *Prima-facie* evaluative judgments would be conditional judgments of the following form:

(a) Insofar as A is more pleasant than B, A is better than B.

All-things-considered judgments are also conditional judgments of the same form, but it is conditioned on all the relevant considerations as follows:

(b) Insofar as all the relevant considerations are considered, A is better than B.

According to Davidson, the Principle of Continence is a rational requirement; the Principle of Continence states that one should always act according to his all-things-considered judgment. ((Davidson 1980), p. 41). The weak-willed agent is irrational exactly by violating the Principle of Continence. Davidson also asserts that by acting against her all-things-considered judgment, by choosing B over A, while judging that A is better than B all-things-considered, the weak-willed agent accepts an *unconditional*, or "all-out" judgment of the form:

(c) B is better than A.

It is according to Davidson the rational conflict between (b) and (c) that makes actions that violate the Principle of Continence irrational. Many philosophers reject Davidson's view that the weak-willed agent accepts (c). In fact, some philosophers think that some forms of *akrasia* involve the agent acting against her "all-out" judgment (See, for instance, (Pears 1982), and (Bratman 1979)). However, until

recently there was wide agreement that weakness of will involved at least acting against one's all-things-considered judgment, or at least against something that was classified as one's "best" evaluative judgment. However, there is no longer consensus among all philosophers even on this point. Holton ((Holton 1999) and (Holton 2004)) and MacIntyre ((McIntyre 1990)) have defended the claim that *akrasia*, or weakness of will, consists in certain types of failures to act on a future-directed intention. I am weak-willed if I form at t_0 an intention to do A at t_1, and yet I do not A at t_1. Of course, this claim needs to be qualified; I might have, for instance, overwhelming reasons to change my mind between t_0 and t_1, in which case, my failure to act as I intended would not count as an instance of weakness of will. More particularly, Holton claims that weakness of will are instances of reconsidering intentions that are "contrary inclination defeating"; intentions that are formed at least partly as "an attempt to overcome contrary desires that one believes one will have when the time comes to act" ((Holton 1999), p. 250).

It is worth noting that these authors will often admit that there is a distinctive failure of rationality that involves acting against one's best judgment (and even concede the label "*akrasia*" to this form of irrationality). Although Holton, for instance, insists that his view captures the "ordinary" use of "weakness of will" by non-philosophers, there are reasons to doubt that this is true; Holton reports only anecdotal evidence in support of this view, and more systematic attempts to test Holton's hypothesis about ordinary usage do not confirm his view (See (Mele 2009)). Holton's view has the advantage of singling out a phenomenon whose

existence there is very little room to doubt; it would be hard to deny that we often fail to act in accordance with our future-directed intentions. On the other hand, the more particular definition of weakness of will as those cases in which the failure concerns a contrary inclination defeating intention does postulate psychological phenomena that might not be as ubiquitous as Holton supposes (I come back to this point momentarily).

More importantly, it is far from clear that weakness of will so defined is a form of irrationality; it is far from clear that reconsidering one's intention is ever irrational *per se*. (see (Broome 2001), and (Tenenbaum)). Let us look more precisely at what Holton considers to be cases of weakness of will. According to Holton, sometimes we form intentions so as to overcome inclinations not to act in a certain way. So I might form an intention not to eat dessert with the purpose as resisting my momentary preferences for certain sweets when they are served right in front of me. According to Holton, if I now revise this intention due to my inclination to eat a certain dessert I exhibit weakness of will; weakness of will is a tendency to revise intentions formed to the purpose of defeating contrary inclination.

However, am I necessarily irrational in eating the dessert, and if so, am I irrational precisely because I have revised my intention? Let us look at two cases in which this kind of intention revision does not seem to be irrational. Suppose that we have the same basic case; namely, an agent who forms an intention not to eat dessert in order to defeat inclinations for sweets. But let us assume now that this is an agent whose modest appetites and particular physiology are such that his health or figure would not be negatively affected even if he were to eat all the desserts that

he would ever feel like eating. However the agent suffers from anorexic tendencies and consequently now forms the intention not to eat any more desserts. Suppose that the agent is served with dessert and, rather than simply turning it down, he reconsiders his intention and comes to the conclusion that he should enjoy himself more, and not be such a "slave of the scale". If the agent decides to eat the dessert on these grounds, there seems to be no reason to impute any kind of irrationality or weakness of will to him. Why should an agent who revises his intention correctly be deemed irrational or weak-willed on the basis of an ill-considered intention he had made in the past? This is not to deny that, *in some cases*, such a decision might express some kind of irrationality. Suppose the agent had come to the conclusion that he does not deliberate well in situations in which he is faced with certain temptations; he concludes, for instance, that he is likely to engage in rationalizations and overlook important features of his choice situation in these circumstances. Given these tendencies, he judges that he ought not to rely on his "momentary" deliberation, but rather stick to his prior intentions. If he now considers revising his intention while retaining (or unwarrantedly revising) the judgment that, all-things-considered, he should not be engaging in (or at least acting in accordance with) such a deliberation, he does exhibit weakness of will, but this is a case that falls straight within the purview of the traditional conception of weakness of will. In this case, even though he is not acting against his better judgment that he should eat the dessert, he is acting against his better judgment that he should not be acting in accordance with his momentary deliberation (or that he should not revise his

judgment in the face of temptation). But the failure of the will is still a case in which the agent does not follow his better judgment.

Finally, the very idea that we form contrary inclination defeating intentions in such a widespread manner would need defense. It is true that in many cases of weakness of will are cases in which we act against a general intention that is supposed to apply to various situations in which we face temptation. But are such intentions "expressly made in order to get over one's later reluctance to act"? Let us take a case that seems to fit well the idea that some of my intentions are formed at least in part in order to defeat contrary inclinations. So perhaps when I form an intention never to drink again in light of my past difficulties with alcohol, one might claim that I am forming the intention precisely to combat the temptation that a cocktail holds for me. If I now find myself in a party and I decide that it would be fine, just this once, to have a beer, I might be indeed manifesting weakness of will. However if we look more closely at this case, it is not clear that the intention would have been formed as "an attempt to overcome contrary desires". Suppose I formed this intention as follows: I used to think that I could drink socially. But now I notice that this is impossible for me; when I start drinking socially, I quickly slip into my old drinking habits. Consequently, I judge that I should never drink; I judge that it is best that I simply refrain from drinking on all occasions (See (Rachlin 2000)). Suppose I now form the intention simply on the basis of this judgment, but with no further aim to overcome contrary desire (of course, I realize that I'll have contrary

desires, but certainly not all intentions formed in the awareness that one will also have desires to act differently are formed as an attempt to overcome contrary desires). It seems that if I fail to act on this intention, I suffer from the exact same kind of irrationality as if I had formed the intention in an attempt to overcome contrary inclination.

And I see no reason to think *a priori* that most cases of my forming such general intentions are cases in which they are also contrary inclination defeating intentions. In fact, even if I don't form the relevant intention, but only make the judgment that the best thing to do is never to drink alcohol, I would still be suffering from the same kind of irrationality; and this is exactly what the traditional conception of weakness of will predicts. In sum, even though Holton is right to think that many cases of weakness of will are failures to act on general intention, and that many such general intentions are formed when the agent recognizes that it is not best to deliberate on the merits of the particular case rather than the merits of the general policy, he is wrong to think that this presents a challenge to the traditional conception of weakness of will.

Although Davidson took it for granted that a rational agent should always follow his all-things-considered judgment, but some philosophers have disputed this claim (See, for instance, (Arpaly 2000)). Huck Finn seems to have acted better by not following his all-things-considered judgment. Huck Finn seems to have thought that, all-things-considered, he should turn in Jim, the runaway slave, to his owner, since Jim was, on Huck Finn's view the slave owner 's rightful property. However, Huck Finn akratically lets his fondness for Jim prevail, and, presumably acts better

by letting Jim run away. But exactly such cases show the difficulty in establishing that we have a genuine case of *akrasia* at hand. After all, one could argue that despite Huck Finn's pronouncements and musings that indicate he acted against what he *thought* (perhaps confusedly) to be something like an all-things-considered judgment (after all, it's unlikely that Huck Finn put the matter to himself in terms of "all-things-considered" judgments), Huck Finn never judges that it is better to turn Jim in, all-things-considered. Of course, since Huck Finn is a fictional character, it is tempting to think that one could simply stipulate that he does make the all-things-considered judgment. But it is not clear that this is a coherent stipulation. Why should we not say that, given that Huck Finn chose the right action in response to the right reasons, we have no reasonable grounds to assert that Huck Finn still judged that, all-things-considered, he should turn Jim in? We can think that we have various pieces of evidence about Huck Finn that might be relevant: Huck Finn's musings about what he ought to do, his various emotions before and after the action, and his actual behaviour. It is not clear which evidence should be conclusive here.

But this leads to a more general concern about whether we have any grounds to ascribe to an agent an all-things-considered judgment. Why should we ever think that an agent's assent to an all-things-considered judgment would be better evidence of what she has most reason to do than her actual behaviour? Isn't it just as good an explanation of an apparently akratic action that the agent changes her mind at the time of the action and then regrets later having changed her mind in this manner? (for a position roughly along these lines, see (Scott-Kakures 1997)). This might be further confirmed by the psychologist George Ainslie's work on weakness

of will (see (Ainslie 2001)). According to Ainslie, typically, behaviour under the heading of weakness of will involves hyperbolic discounting; we do not simply discount linearly future rewards, but the rate of discounting changes dramatically as we approach a certain reward. So even though Tuesday I might prefer waking up sober the following Monday over drinking a lot on Sunday, I will discount the rewards of being sober more dramatically as Sunday evening nears until I finally experience a preference shift as I walk into the bar. We could then say that an agent in such a situation changes her evaluative judgments in lockstep with her preference shifts and that her claim she chose against her best judgment can be understood as expressing the evaluative judgments that she made before and after the action, but not *at the time of the action* (for an attempt to use Ainslie's work in order to show that there is at least no intentional counterpreferential choice, see (Heath 2008)).

One might object that this line of reasoning establishes at best epistemological difficulties in establishing that we *know* that an agent is akratic in a particular case, but gives us no reason to suspect that *akrasia* might not be a widespread phenomenon, whether or not we can ascertain its existence. However, it is not clear that all-things-considered judgments have any psychological reality that is independent of what one is warranted to ascribe to the agent in attempting to provide intentional explanations of his behaviour. Davidson himself argued that the ascription of mental states such as beliefs and desires depends on a constitutive use of the Principle of Charity: the agent's beliefs and desires are those that would make most sense of her behaviour in the light of the assumption that the agent is a "believer of truths and lover of the good" ((Davidson 1980), p. 222). But even if one

does not accept Davidson's extreme contention, it seems plausible to think that which mental states are correctly ascribed to an agent is partly determined by the role of such mental states in explaining the agent's actions, and that, *ceteris paribus*, we should not attribute to the agent needless irrationality.

But this very thought gives us a path to confirm our confidence that cases of *akrasia* are not only possible but widespread. For it is plausible on many occasions to conclude that failing to ascribe *akrasia* would be a greater violation of the principle of Charity than not ascribing it. Suppose I am a divorce lawyer whose famous client has told me various risqué stories about herself and her husband. It is tempting to gossip about the case to my friends, but I know that I have overwhelming prudential and moral reasons not to break the confidentiality of my client. I understand very well the force of these reasons; I know for instance that my career depends on it, and that it would not be fair to my client to spread around intimate aspects of her life. Moreover, I have on many occasions resisted the temptation to gossip about it. Now I am at a party where people have been gossiping and I finally succumb to the temptation to tell a salacious story my client related to me. After telling the story, I immediately regret what I did and judge that I have done something tremendously stupid. In this case, denying that I behaved akratically would have very implausible consequences about how I revise judgments. The judgment that I should not break my client's confidentiality was, *ex hypothesis*, formed on good grounds and perhaps after a great deal of reflection; it was probably grounded on deep features about my character and central aspects about some of the projects that are very important in my life. This judgment was

also reaffirmed just after my indiscretion. So if indeed I changed my mind momentarily, this would mean that I have revised a well-grounded, previously stable judgment, on the basis of reasons that I am clearly capable of knowing are bad reasons, when no new information is available, and then immediately reverted back to the original judgment despite, again, having no new evidence. This would be highly irrational behaviour; certainly, no less irrational than simply acting against one's all-things-considered judgment. And, again, whatever one thinks about the Principle of Charity, it seems a much simpler explanation to say that the agent acted against her best judgment than to say that the agent underwent all these irrational revision processes.

Notice that the argument that the ascription of *akrasia* was warranted depends at least in part on the claim that the agent formed the evaluative judgment for good reasons, and held it in a stable manner. Thus it seems that *akrasia* is at the very least more easily attributed in the cases in which the agent acts contrary to her *knowledge* of the good, rather than when she against what she merely believes to be good. One of Davidson's many contributions to the literature on *akrasia* has been to claim that *akrasia* is best defined in terms of action contrary to an evaluative *belief* rather than, as it had been traditionally defined, in terms of action contrary to a piece of practical knowledge. Davidson's position has been nearly unanimously accepted in the literature on *akrasia* (for a notable recent exception, see (Engstrom 2009)). However, reflection on the possibility of *akrasia* raises the suspicion that the traditional view of *akrasia* might be the correct one.

Ainslie, G. 2001. *Breakdown of Will*. Cambridge University Press.
Aristotle. 1985. *Nichomachean Ethics*. Indianapolis: Hackett Publishing Co.
Arpaly, N. 2000. "On Acting Rationally against One's Best Judgment*," *Ethics*, **110**: 488-513.
Bratman, Michael. 1979. "Practical Reasoning and Weakness of the Will," *No*$\sqrt{}^{a}$*s*, **13**: 153-71.
Broome, J. 2001. "Are Intentions Reasons? And How Should We Cope with Incommensurable Values?". In Christopher and Ripstein, eds, *Practical Rationality and Preference: Essays for David Gauthier*: Cambridge University Press.
Davidson, Donald. 1980. *Actions, Reasons, and Causes.* New York: Oxford University Press.
———. 1980. "How Is Weakness of the Will Possible?". In Davidson, ed, *Essays on Actions and Events*. Oxford: Clarendon Press.
Engstrom, Stephen. 2009. *The Form of Practical Knowledge*. Cambridge, Mass.: Harvard University Press.
Heath, J. 2008. *Following the Rules: Practical Reasoning and Deontic Constraints*. New York: Oxford University Press.
Holton, R. 1999. "Intention and Weakness of Will," *The Journal of Philosophy*: 241-62.
———. 2004. "Rational Resolve," *The Philosophical Review*.
Kant, I, AW Wood, G Di Giovanni, and RM Adams. 1998. *Religion within the Boundaries of Mere Reason and Other Writings*. Cambridge Univ Pr.
Korsgaard, Christine M. 1997. "The Normativity of Instrumental Reason". In Cullity and Gaut, eds, *Ethics and Practical Reason*. New York: Clarendon Press.
McIntyre, A. 1990. "Is Akratic Action Always Irrational?". In Flanagan and Rorty, eds, *Identity, Character, and Morality*. Cambridge, Mass.: MIT Press.
Mele, A. 2009. "Weakness of Will and Akrasia," *Philosophical Studies*: 1-14.
Pears, D. 1982. "How Easy Is Akrasia?," *Philosophia*, **11**: 33-50.
Rachlin, H. 2000. *The Science of Self-Control*. Harvard University Press.
Scott-Kakures, D. 1997. "Self-Knowledge, Akrasia, and Self-Criticism," *Philosophia*, **25**: 267-95.
Smith, Michael. 1994. "The Moral Problem". *Philosophical theory.* Oxford: Blackwell Publishers.
Stocker, Michael. 1979. "Desiring the Bad: An Essay in Moral Psychology," *Journal of Philosophy*, **76**: 738-53.
Tenenbaum, S. "Intention and Commitment".
Tenenbaum, Sergio. 2007. *Appearances of the Good: An Essay on the Nature of Practical Reason*. New York, Cambridge: Cambridge University Press.