# IS THERE SUCH A THING AS WEAKNESS OF THE WILL?

Kathrin Glüer

## *1. Introduction*

In this paper, I shall defend a version of skepticism about weakness of the will. To put it in a nutshell, the skepticism I find attractive claims that the paradigmatic cases of putative weak-willed action do not involve any particular form of *practical* irrationality at all. Those that cannot be completely explained away are better understood as cases of *self-deception*.

Weakness of the will is a topic that has exercised philosophers ever since antiquity. In this ongoing debate, the issue has never simply been how to understand and account for weak-willed action; rather, the very existence of the phenomenon has always been at stake at the same time. Weak-willed action generates problems for certain philosophical accounts of action explanation whose principles seem to exclude its very possibility. Unlike with other phenomena, however, this might not simply show that these philosophical accounts are false. To many, our pre-theoretic grasp of the phenomena in question does not seem strong enough to warrant such a conclusion. For in order even to generate conflicts with philosophical accounts of action explanation, precise descriptions of weak-willed action need to be formulated, and these might easily seem so far removed from everyday ascriptions of weakness that the intuitive evidence is rather weak. Faced with such conflicts, all of the following options thus seem open: give up the account of action explanation that generates the problem, embrace skepticism about weakness of the will, or modify the description of weakness that generated the conflict.

However, the link between philosophical concepts and pre-theoretic conceptions of weakness actually seems even more strained than this picture suggests. Both the descriptions of weakness, and the action-theoretic principles it conflicts with, are, after all, supposed to figure in accounts of our actual practices of action explanation. *All* of them thus appear motivated, to some degree, by our pre-theoretic conceptions and practices. Therefore, if there is conflict on the more theoretical level, this might well reflect opposing tendencies and less than wholly consistent practices of everyday action explanation. And it seems to me that this is exactly what it does.

On the level of philosophical theory, "weak-willed" actions, at least on one traditional understanding, owe their (misleading) description to a basic confusion about the relevant notion of evaluation. The notion of evaluation traditionally used to define weakness of the will admits of two kinds of readings, normative and explanatory ones. Read normatively, denying the possibility of weakness appears counterintuitive, while weakness at the same time seems to be defined in terms of an explanatory notion of evaluation. This to some degree reflects two opposed tendencies operative on the level of everyday action explanation and evaluation: On the one hand, we do tend to regard ourselves and our fellow mortals as good-willed but weak and easily falling for temptation; on the other hand, however, we also tend to take at least *their* falling to show something about who they really are and what they really want.

Given all this, it is not surprising that despite a substantial literature on the topic there is today no substantial consensus about what weakness of the will is. A general case for skepticism would either have to show that there is no such thing under any of its suggested definitions or that one of them is the correct definition. I have no idea how to do either. In this paper, I shall therefore offer a limited defense of the particular version of skepticism about weakness of the will I find appealing: First, I shall identify the more traditional approach to weakness of the will that takes a weak-willed action to violate a so-called *better judgment* as our target. More precisely, I shall focus on one particular, very influential proposal for integrating the possibility of akratic action into a general theory of action explanation: the proposal Donald Davidson put forward first in his by now classical 1970 paper "How is weakness of the will possible?". After lining out this proposal in some detail, I shall try to convince you that this very proposal, if looked at long and hard, dissolves into exactly the kind of skepticism sketched above. I shall also argue that this conclusion holds for any account of action explanation that shares certain crucial features with the Davidsonian.

## *2. What is weakness of the will?*

The notion of *weakness of the will*, *akrasia* or *incontinence* - I shall use all three terms, and I shall use them interchangeably - applies both to particular actions and to a character trait. Here, I shall be solely concerned with akratic actions. What seems clear is that an acratic action is a *free, intentional action*. And it is an *irrational* action in some sense to be specified. On the understanding of akrasia to be considered here, acratic action is irrational in that it violates a certain kind of practical judgment. Let's look at some examples.

My first example is taken from Ariela Lazar:

> At a party, you offer me a cigarette and I consider your offer. On [the] one hand, I reckon that smoking a cigarette now would be very enjoyable. On the other hand, I believe that smoking is detrimental to my health. Not only would smoking this cigarette be harmful in itself, I think, but it would also increase the probability of my picking up an abandoned habit of smoking. In short, accepting your offer for a cigarette decreases my chances of leading a long and healthy life. I consider this goal superior to that of enjoying the short-term pleasure of smoking. Given my own values and goals, as well as the extent to which I believe accepting your offer would interfere with or satisfy then, I judge that I ought to decline your offer. But then, I take the cigarette, muttering uncomfortably: 'Just this once' (Lazar 1999, 381).

From this first and rather typical example, we can gather the following: Weak-willed action is irrational in some *internal* or *subjective* sense. It violates one's own values and goals and how one believes one best acts in order to pursue them. It does not matter whether these beliefs are true or whether these goals are truly worthy of pursuing. What matters is that they are one's own and that one does not act accordingly.

In this example, the consideration is between short-term pleasure and long-term interest. That is fairly typical for the examples found in the literature. The other traditional struggle is that between reason and passion. But whether an action is weak-willed is not determined by which of these wins. Reason might win, or long-term interest might win, and the action might still be weak-willed. For an action is weak-willed as soon as it goes against what the agent himself values most. And he might be the person who doesn't give a damn about long-term health considerations. He might value immediate sensual pleasure much higher. Neither is weak-willed action necessarily immoral or selfish. Davidson nicely brings out this point in the following, intentionally atypical example:

> I have just relaxed in bed after a hard day when it occurs to me that I have not brushed my teeth. Concern for my health bids me rise and brush; sensual indulgence suggests I forget my teeth for once. I weigh the alternatives in the light of the reasons: on the one hand, my teeth are strong, and at my age decay is slow. It won't matter much if I don't brush them. On the other hand, if I get up, it will spoil my calm and may result in a bad night's sleep. Everything considered I judge I would do better to stay in bed. Yet my feeling that I ought to brush my teeth is too strong for me: wearily I leave my bed and brush my teeth (Davidson 1970: 30).

In both these examples, the agent does something that goes against the result of some process of deliberation or practical reasoning. By means of that, the agent reaches a verdict about what it is best for him to do, or at least a verdict about which of two options *a* and *b* is better. He comes to the conclusion that *a* is better than *b* – nevertheless, he performs *b*. And this is not the result of any

reconsideration; on the contrary, he holds the better judgment at the very time of action. The irrationality of akratic action is here determined as purely *practical* in character: It's not that our agent makes some mistake in his reasoning, or forgets or revises the result; rather, he acts against an internally correct practical judgment that he holds at the time of action.[1]

This traditional understanding of acratic action as free, intentional action against one's own better judgment is what Davidson tries to capture in his definition:

> D. In doing *x* an agent acts incontinently if and only if: (a) the agent does *x* intentionally; (b) the agent believes there is an alternative action *y* open to him; and (c) the agent judges that, all things considered, it would be better to do *y* than to do *x* (Davidson 1970, 22).

And it is Davidson's attempt to make room for this kind of akrasia in his account of action explanation that I shall engage with in the rest of this paper.

[1] As indicated above, this is not the only suggestion how to understand weakness of the will on the market. As Rorty has pointed out, there are various places in the chain of reasoning leading to action where an "akratic break" can occur (cf. Rorty 1983). Furthermore, akrasia can be characterized synchronically or diachronically. Thus, it has recently been argued that the condition of acting against a better judgment is not necessary for akrasia. An agent might act in accordance with his better judgment and still act weak-willed. Examples supposed to illustrate this involve a prior decision to act against a better judgment that is not acted on when the time comes. The reason for abstaining, however, is not the (still held) better judgment but something else, too much fear for instance (cf. Mele 1987, 54. See also Jackson 1984, 4; Mele 1994, 282). However, such actions are supposed to be weak-willed not because they are done for a reason *other than* the better judgment, but because they are not in line with a prior decision. The weakness is located in the *change* of mind. This is, therefore, a different phenomenon from the one described in the examples above. Both kinds of descriptions undoubtedly have some root in ordinary ascriptions of weakness but they should be kept apart on a more theoretical level.

Another influential suggestion is to spell out weakness of the will in terms of the relation between first and *second order desires*. Roughly, the idea is that in weak-willed action one acts on a first order desire that one has a second order desire not to have (cf. Frankfurt 1971, Jeffrey 1974, Schiffer 1976). It seems to me that the very notion of a second order desire is more problematic than these discussions presuppose. One of the questions here concerns how such desires are to figure in processes of deliberation or practical reasoning. If a second order desire simply were a reason on a par with any other reason, there would not necessarily be any irrationality in overruling it. But if, instead, irrationality is generally determined from the level above, this characterization of akrasia seems threatened by vicious regress. For the irrationality of acting (or desiring) against a desire from the level above would depend on there being yet another desire two levels above determining this. And so on (cf. Peacocke 1985, 54).

*3. Davidson on explaining intentional action*

Davidson claims that it is self-evident that weakness of the will, as defined in D, exists (cf. Davidson 1979, 23). His problem is that the principles of action explanation that seem to make it impossible are at least equally self-evident. The conflict we are interested in arises, or would seem to arise, between clauses (a) and (c) of Davidson's own definition. For being an intentional action *is* being done for a reason, *is* having a certain kind of explanation in terms of the agent's reasons the very existence of which seems to be denied by condition (c).

Davidson has a number of papers that are concerned with weakness of the will. All the later ones, however, treat weakness of the will as just one kind of irrationality out of many. The bigger problem in which these later discussions figure is that of reconciling a Davidsonion picture of what understanding intentional states and actions consists in with the existence of irrationality. For on a Davidsonian model, a very basic kind of rationality is a condition on the possibility of having any intentional states at all. In such a setting, it might well seem as if there was something paradoxical about irrationality. Davidson puts it this way: "The paradoxical consequence is that explaining irrationality necessarily employs a form of explanation which rationalizes what it explains" (Davidson 1985, 347). The danger is, it seems, that we are after some explanation that will make the very phenomenon we try to explain disappear.

The first question to be asked here is what is meant by "explaining irrationality". As far as I can see, it actually means two different things. For one, we need to explain how there can be irrationality at all. In this sense, it simply means that our account of the intentional must not make irrationality impossible. We must be able to acknowledge its existence. If our account is able to tolerate irrationality, we have, in this sense, explained how it is possible to be irrational.

However, there is another sense of explanation in play here, especially when Davidson says that "explaining irrationality necessarily employs a form of explanation which rationalizes what it explains". For any irrational action, that is, there has to be an explanation that rationalizes it. This is because what we are after is irrational action, not non-rational events or happenings. Such actions are intentional and that means, according to Davidson, that they have been done for reasons. Surely not necessarily for the best reasons, and in case of weak-willed action, not even the best reasons available to the agent himself, but there has to be a reason for which they were done.

The most basic such reasons Davidson calls "primary reasons". For any intentional action, there has to be a primary reason. It consists of a belief-desire pair related to the action in specific ways. Davidson explains: "What is to be

explained is the action, say taking exercise. At the minimum, the explanation calls on two factors: a value, goal, want or attitude of the agent, and a belief that by acting in the way to be explained he can promote the relevant value or goal, or will be acting in accord with his attitude" (Davidson 1982, 293). A minimal explanation of Alfred's going for a run this morning thus consists in his desire for a run. A little more technically, any kind of motivating state that can enter into a reasons-explanation is called a desire or, even more technically, a pro-attitude. According to Davidson, desires are directed at actions of a certain type. In this case, Alfred had a desire for actions of the type going-for-a-run. To explain his particular run this morning, we also need a belief, namely that this particular action is of the desired type.

In order to explain Alfred's run, belief-desire pair and action must be related in two very different ways: "First, there must be a logical relation. Beliefs and desires have a content, and these contents must be such as to imply that there is something valuable or desirable about the action. Thus a man who finds something desirable in health, and believes that exercise will make him healthy can conclude that there is something desirable in exercise, which may explain why he takes exercise" (ibid.). The main idea of a reasons-explanation is that it explains by showing what was desirable about the action from the agent's own point of view. They rationalize the action by giving reasons from the agent's own perspective, by showing why he wanted to do what he did. Such an explanation necessarily exists for every intentional action; if it is not rationalizable in this minimal sense, if it cannot be understood why it was done even from the agent's own point of view, then there is no reason to think it's an intentional action at all.

Of what kind is this logical relation Davidson talks about? It cannot be a simple deductive or syllogistic one, he argues. It cannot be a simple Aristotelian practical syllogism where the intention or the action are deductively implied by the premises, the desire and belief, that is. There are two reasons for this. One is that an agent might have *more than one reason for a particular action*. Alfred, for example, has a desire for doing something healthy, but he also wants to get away from his desk for a while and to show off in his new running outfit. For which of these reasons did he go running? It might be all of them together, of course, but it might just as well be only some, or even only a single one of them. If, however, reasons and actions were related deductively, so that from having a reason an action would automatically follow, it would not be possible to distinguish *the* reason an agent acted for from other reasons he had.

What's more, there would be no way of handling *conflicting reasons*. Alfred might, after all, have reasons for and against his run. It might be raining and he might desire not to get wet, for instance. If this reason were deductively related to his not running, and at the same time he had the desire to do something for his health by running, Alfred would end up in a straight contradiction. If we conceive of the conclusion of a practical syllogism as an action, he would both run and not run. If we think the conclusion is an intention he would intend both to run and not to run.

Therefore, the logical relation between primary reasons and actions needs to be loosened, so to speak. There needs to be some sort of a logical gap between the reasons and the actions that allows for reasons an agent has for and against a certain action without them being the reasons on which he acted. This is why Davidson says: "Beliefs and desires have a content, and these contents must be such as to imply that there is *something valuable or desirable about the action*" (ibid.). Beliefs and desires do not imply the action, or the intention, but they do imply that from the point of view of the agent there is something valuable or desirable about the action. This relation between reasons and action can instructively be compared to the relation between a belief and the evidence we have for it.

However, in giving reasons-explanation we aim at providing *the* reason the agent acted for. Being *a* reason, in other words, is only necessary, not sufficient for being *the* reason. What this shows, Davidson argues, is that reasons-explanation are *a species of causal explanation*. Primary reason and action do not only have to stand in a logical relation, they also have to be related as cause and effect: "Second, the reasons an agent has for acting must, if they are to explain the action, be the reasons on which he acted; the reasons must have played a causal role in the occurrence of the action" (ibid.).

Both of these are necessary conditions on reasons-explanations, Davidson holds, and (primary) reasons themselves are necessary for intentional action. Even an irrational action has a primary reason that explains it. And that's what's paradoxical about irrationality: "This much of the analysis of action makes clear why all intentional actions, whether or not they are in some further sense irrational, have a rational element at the core; it is this that makes for one of the paradoxes of irrationality" (ibid.). The trick, then, is to explain what the further irrationality consists in. And that brings us back to akratic action.

*4. How is weakness of the will possible?*

Weak-willed action results from an internal conflict of reasons. It is clearly done for *a* reason, for it is done for *one* of the reasons in conflict with each other. The problem is not that there is *no* reason that would entail that there is *something* desirable about the action; rather, it is that there are other reasons that entail that there is something *even more desirable* about not performing the action. Davidson puts it like this: "The difficulty is that their existence [i.e. the existence of akratic actions] challenges another doctrine that has an air of self-evidence: that, in so far as a person acts intentionally he acts, as Aquinas puts it, in the light of some imagined good. This view does not, as it stands, directly contradict the claim that there are incontinent actions. But it is hard to deny that the considerations that recommend this view recommend also a relativized version: *in so far as a person acts intentionally he acts in the light of what he imagines (judges) to be the better*" (Davidson 1970, 22, emph. mine).

Think of Alfred and his run again. He has, we said, reasons for and against the run. What he had to decide is whether to run or not to run. He runs. What Davidsons wants us to see is that we can *only* explain that in terms of his reasons if we say that Alfred thought that his reasons for running were *better* than his reasons against. We have, that is, the kind of reasons-explanation required for intentional action only if the action is in accordance with, and motivated by, what we have called a *better judgment* above. And that is why, even though they clearly are performed for reasons, akratic actions might seem to be impossible: Nothing can be intentional and against a better judgment at the same time.

Davidson formulates the underlying assumptions in principles P1 and P2:

> P1. If an agent wants to do *x* more than he wants to do *y* and he believes himself free to do either *x* or *y*, then he will intentionally do *x* if he does either *x* or *y* intentionally. (…)
>
> P2. If an agent judges that it would be better to do *x* than to do *y*, then he wants to do *x* more than he wants to do *y* (Davidson 1970, 23).

But from P1 and P2 it follows that if an agent judges that it would be better do *x* than to do *y*, and he believes himself free to do either *x* or *y*, then he will intentionally do *x* if he does either *x* or *y* intentionally. Therefore, incontinent action, as spelled out by Davidson, seems to be excluded. P1 and P2 seem inconsistent with the claim of its existence.

But, or so Davidson claims, they are not. P1 and P2 in fact hold for all intentional actions, even akratic ones. The difference this compatibility claim turns on is the difference between a judgment "that it would be better to do *x* than to do *y*", as in P2, and a judgment "that, all things considered, it would be

better to do *y* than to do *x*", as in D. These judgments do not contradict each other. And that is what allows Davidson to allow for akratic action.

To see why, we have to look somewhat closer at how his model of practical reasoning deals with conflicting reasons. In order to allow for conflict at all, we saw above, having a certain desire cannot be represented as judging that actions of a certain type are desirable. Rather, desires are to be represented as judgments that having a certain feature counts in favour of an action, while having other features counts against it. Such judgments Davidson calls *prima facie judgments*. The 'prima facie' operator here does not operate on predicates of actions or on single sentences, but is a sentential connective. Davidson's example: "That an act is a lie prima facie makes it wrong" (Davidson 1970, 38). The trick is that this way, you cannot detach; "*a* is wrong" cannot be derived from "*a* is a lie" and the prima facie judgment. No 'unconditional' judgments, that is, follow from prima facie judgements. Therefore, considerations that speak in favor of *a* can be conjoined with considerations that speak against it without contradiction. Moreover, this construction allows for a process of weighing reasons against each other that arrives at what Davidson calls an "all-things considered judgment".[2] Even this is a 'conditional' judgment, but what it is conditional on would, ideally, be all relevant reasons the agent has. Such judgments would be "All things considered, I should do *x*" or "All things considered, *x* is the best thing to do" or "All things considered, *x* is better than *y*". Let's stick with the last to avoid unneccessary complications.

On a Davidsonian model of weighing reasons, desires are represented as prima facie judgments. Weighing prima facie judgments can only result in judgments that are themselves of prima facie character. To form an intention, however, is to reach a conclusion as to what is better, period. To form an intention, that is, you have to detach. You have to take the step from the result of the practical reasoning, from the all-things-considered judgment, to what Davidson calls an all-out judgment, a judgment of the form "*x* is better than *y*". This step, however, is not logically warranted; the logical gap that was introduced into the relation between reasons and action to allow for conflict is, according to Davidson, not closed by the modified model of practical reasoning. The strongest reason does not have to be the strongest cause.

[2] This process of weighing can be described with the help of decision theoretic means, that is, by means of the calculus of maximising utility. "The idea is simple: the value an agent puts on a contemplated action depends on the values he places on the various ways he thinks the world may turn out to be given the action, and how probable he thinks the various outcomes are" (Davidson 1985b, 199). The idea is spelled out in detail in Jackson 1984.

It is this logical gap that Davidson utilizes in allowing for incontinent action. Here, the agent reaches an all-things-considered judgment to the effect that *y* is better than *x*. This judgment is a 'conditional' prima facie judgment, however. So there is nothing that logically prevents the agent from nevertheless forming an 'unconditional' or all-out judgment to the effect that *x* is better than *y*. This corresponds to his intention to *x*. Therefore, the akratic agent not only has *a* reason for his action, he also judges, wants, and acts in accordance with P1 and P2.

But then: what's irrational about incontinent action? According to Davidson, there is a principle constitutive of practical rationality that does, after all, link all-things-considered judgments and all-out judgments or intentions. He calls it the "principle of continence" and it says: "perform the action judged best on the basis of all available relevant reasons" (Davidson 1970, 41). This principle is so fundamental to intentional action that a creature whose behaviour continually did not accord with it could not be counted intentional. At most, occasional violations are tolerable, if behaviour is to be understandable as intentional action. Occasional violations, however, do not destroy interpretability. And thus, occasional incontinence is possible.

*5. Motivation and evaluation*

In the last two sections I would first like to defend the Davidsonian account against one of its most popular criticisms. Then, however, I shall argue that this very line of defense ultimately turns weakness of the will into self-deception.

Davidson's account of akrasia has dominated the literature for 33 years now. While it is rarely embraced, there is hardly an article on the topic that does not position itself in relation to it. As far as I can see, it is especially principle P2 that has received tremendous resistance, resistance that, for the most part, seems misguided to me. The line of criticism I would like to focus on in this section makes use of the idea of a possible *divergence between motivation and evaluation* in one or the other form. P2 does not allow for that. And this 'internalism' appears to many at least far from self-evident, if not outrightly counterintuitive.[3]

To see why such criticism, pointing to the seemingly obvious possibility of wanting diverging from judging best, is beside the point, we should ask what it exactly is that principles P1 and P2 are supposed to do: These principles are meant to connect a rather austere model of action *explanation* with the ordinary concept of wanting. The model works with beliefs, desires and intentions. Since it models decisions to act as more or less explicit processes of practical rea-

[3] Cf. Watson 1977, Pears 1984.

soning, it *represents* all of these states in terms of judgments. Desires or pro-attitudes, that is, motivational states of every kind, are represented as evaluative judgments. This choice is motivated by considerations like the following:

> [S]omeone who says honestly 'It is desirable that I stop smoking' has some pro attitude towards his stopping smoking. He feels some inclination to do it; in fact he will do it if nothing stands in the way, he knows how, and he has no contrary values or desires. Given this assumption, it is reasonable to generalize: if explicit value judgements represent pro attitudes, all pro attitudes may be expressed by value judgements that are at least implicit (Davidson 1978, 86).

The decision to represent desires as evaluative judgements in practical reasoning, however plausible in the light of such considerations, remains something of a "stipulation" (ibid.). Ultimately, it is by stipulation that evaluation and motivation cannot be divorced in a Davidsonian model of practical reasoning. The value judgments figuring in this model *express* or *represent* desires, that is, all kinds of pro-attitudes or motivational states an agents has or is in, whether he explicitly thinks about them or not. P1 and P2 do nothing but claim that the desire with the highest value, the desire with the strongest motivating force, is the one an agent acts on. And that does in fact seem pretty self-evident.

The notion of evaluation used here is, consequently, a rather thin one.[4] It is thin or "weak" in two different senses: On the one hand, it is completely *internal*. It is the agent's own desires and values, and only those, that are expressed and options are evaluated as better or worse in their light and their light only. On the other hand, this notion of evaluation is completely *general* or *topic-neutral*. Options are evaluated as better or worse simpliciter, not as morally better or prudentially better or more rational in any strong sense.[5] Such a concept of evaluation, it seems to me, thinned out as it is, would vanish completely if we divorced it from motivational force. Motivational force *is* what all the pro-attitudes have in common, if anything. More importantly, however, its very thinness or weakness is, at the same time, the strength of this notion. All its usefulness derives from being thin in the two senses just spelled out. For it is what gives this model of practical reasoning its *explanatory power*. Topic-neutrality allows for weighing, and, thus, for a logical (or at least rational) connection between different and maybe even conflicting reasons and action.

[4] David Pears has called it a "weak" or "decision theoretic" concept of evaluation (cf. Pears 1984).

[5] The 'topics' a judgment employing this topic-neutral concept is meant to be neutral about thus are not descriptive kinds of action, but *kinds of evaluation* that could be applied to actions

Together with internality, those reasons being the agent's own, topic-neutrality thus allows for explaining why *he* did what he did.

This concept of evaluation, it thus seems to me, is, by its very nature – be that ever so artificial – *explanatory or descriptive*. And that means that *if* the values and beliefs we feed into our model of practical reasoning, that is, the pro-attitudes of an agent, their "strengths" expressed in terms of numerical values, and the conditional probabilities of their outcomes, *are those an agent actually has*, then he *will* do whatever the reasoning grinds out as the better, or even the best, option. Unless, of course, something interferes or he changes his mind. And this seems to be exactly what we would expect from an explanation: He did what he did *because* he had those reasons. It is very hard to see how he could have done otherwise, for if he could, how would having these reasons explain what he does?

This does not mean, of course, that models of practical reasoning cannot be used to do something *normative*. Most decision theoretic uses of such models are, in some sense, normative. For practical reasoning to be normative, however, it needs to be possible that *what an agent actually does diverges from what he should do or ought to do*. Given that basic belief-desire models contain only one numerical parameter attached to pro-attitudes, such divergence cannot be captured *in* the model; at most, the model can be used to model practical reasoning resulting in one or the other, but not both at the same time. If the model is used normatively, the divergence will show as a divergence between its outcome and what an agent actually does. To capture reasoning resulting in what an agent should do, as opposed to what he will do, the concept of evaluation used in the model has to be *'thickened*', however. There are, it seems to me, two basic ways of doing this. They correspond to the two senses in which the explanatory concept of evaluation is 'thin': It is thin, we said, first, in that it is completely internal. The values it calculates with are those the agent actually has. In this sense, reasoning is captured from the agent's own perspective. Even from an agent's own perspective, however, what he judges he should do might diverge from what he actually does. We can model this if we read the concept of evaluation in a *topic-restricted* or *topical* sense, for instance, as moral evaluation. This way, any kind of particular *should* or *ought* can be opposed to any kind of equally topical *wanting*, and any kind of topical should can diverge from what the agent topic-neutrally wants and, therefore, will do.[6]

Secondly, the explanatory concept of evaluation is thin in exactly this sense, that it is topic-neutral. The reasoning results in a better judgment that has taken

[6] Will do, that is, if he does not change his mind and nothing interferes.

into account everything that speaks for or against an action in any sense whatsoever. Even a topic-neutral concept of evaluation can be used normatively, it seems to me, but then, we need to bring some *external standard* to bear on *which attitudes* an agent *should* have and *what strength* he *should* attach to them. We need to give up the complete internality of the values calculated, that is. This is, I believe, what goes on in normative decision theory; there, it is not only assumed that a rational agent is internally rational, but also that a rational agent has certain values and not others.[7]

Where does all this leave us with respect to Davidson's P2? Principles P1 and P2, we said above, serve to connect the ordinary concept of wanting with the belief-desire account of action explanation. This account is non-normative; the concept of evaluation employed in its better judgments is completely internal and topic-neutral. Therefore, P2 does not allow for any divergence of evaluation and motivation. And this is no accident; it is non-normative reasons-explanations that are constitutive of intentional action, according to Davidson, and his very problem was that this constitutivity seems to make akratic action impossible. That is, the kind of explanation required for any intentional action is explanation in terms of topic-neutral, internal evaluation. And it is in terms of this very concept of evaluation akrasia itself is understood. Akrasia is *defined* as a kind of practical irrationality where what is violated is a completely internal and topic-neutral evaluation. Complaining that this does not take the possibility of motivation diverging from evaluation into account thus is simply beside the point.

### *6. Akrasia and self-deception*

If we pursue this line of defense of the Davidsonian principles, however, it might well seem as if it was not the possibility of akrasia, but skepticism about it, that followed. The concept of evaluation employed in all the better judgments Davidson's account makes use of, is, after all, the same. It has to be, for how could there otherwise be a rational connection between practical reasoning in terms of prima facie judgments and intentions (all-out judgments)? The concept of evaluation employed in the all-things-considered judgment that according to Davidson results from practical reasoning thus is the same old internal, topic-neutral concept that is also employed in the intention. It's character thus spreads to the whole model. We seem, therefore, forced to the conclusion that if the reasons an agent has do result in the conclusion that *y* is better than *x*, then he *will* do what he judges is better.

[7] The two kinds of thickness can, of course, also be combined.

Realizing that this claim results from consistently employing a purely explanatory concept of evaluation should take away much of its apparent counterintuitiveness. Most of the paradigmatic examples of akratic action do, after all, involve conflicts between different normative, or at least different topic-restricted, evaluations. Let's look at the smoking case desribed above: There are basically two possibilities of viewing it in the light of these considerations. The first of them makes use of the idea that the case is *mis- or at least under-described*. That is, the action here does not violate a completely internal and non-particular evaluation; what is really acted against actually is an evaluation employing a 'thicker' concept of evaluation. As this is a long-term interest vs. short-term pleasure case, the 'thicker' concept probably is that of prudential evaluation. Taking into account how the process of deliberation preceding the action is described in the example (all relevant reasons seem to have been considered), the concept appears to have been 'thickened' by attaching values to prudential desires that were higher (relative to other desires) than their actual motivational force. What was deliberated, in other words, was *what a prudentially more rational person would do*. If this is the correct 'diagnosis', however, the case was misdescribed. It is, after all, not a case of acting against a completely internal all-things-considered judgment and, therefore, not a case of akrasia at all.

The other possibility comes into play in case our agent digs her heels in when confronted with this 'diagnosis'. She denies that her better judgment was non-internal (or topical) in any sense. In that case, we will have to say that she was *self-deceived*. She did attach some values to her desires that did not correspond to their motivational force but not because she was reasoning with a 'thicker' concept of evaluation. Rather, she *believed* that those were the actual values. She was *mistaken* about what she valued how much, at least at this particular time,[8] and therefore her judgment diverged from what she actually did. That, however, is not such a radical suggestion as it might seem. For what she is mis-taken about is what she really, honestly, deep down wants most. That's probably the most common kind of self-deception there is. We often think that our values are such and such, that we are the kind of person who values the good and the

[8] Adopting this line does not commit us to deny that, in some sense, she nevertheless valued not smoking more than smoking. For what we are talking about here is values attached to desires and pro-attitudes towards (aspects of) *particular actions* at particular times. Out of these 'occasional' values, we might very well in some way construe values in a more 'standing' sense, values that an agent attaches to *kinds of actions over time*. Values in this standing sense do not require acting (or reasoning) according to them on every particular occasion, even though they do not allow for violation all the time. This notion of value is, it seems to me, pretty close to everyday talk about what someone values.

tasteful and that even now, on this very occasion, we would rather go and see the modern art exhibition than remain on the sofa reading a kitschy novel. But that is often wishful thinking. What we actually do, namely remain on the sofa, should tell us something about ourselves, it seems to me. Not necessarily that we are a kind of person who does not value the good and the sophisticated,[9] but at least that right now, we value inertia more. The irrationality involved in such self-deception, however, is *not practical* in character. On the contrary, given what the self-deceived agent's desires really are, what she does is perfectly rational. A case of such self-deception, therefore, is not a case of akrasia either.[10]

Davidson himself does, of course, not draw these conclusions, even though there are passages that point in this direction. Thus, he says, for instance:

> Why would anyone ever perform an action when he thought that, everything considered, another action would be better? If this is a request for a psychological explanation, then the answers will no doubt refer to the interesting phenomena familiar from most discussions of incontinence: self-deception, over-powering desires, lack of imagination, and the rest (Davidson 1970, 42).

This is a bit puzzling. For if, as we just saw, the 'psychological explanation' of an akratic action involves self-deception that means that *irrational beliefs* were involved in the deliberation leading to the violated better judgment. Beliefs, that is, that *misrepresented* the agent's desires and their strengths. What is supposed to explain intentional action, however, is the first-order (beliefs and) *desires an agent actually has*, not *beliefs about them*. The gap Davidson feels forced to introduce between prima facie reasoning and intending in order to allow for akrasia thus would seem misplaced.[11] Let me explain:

---

[9] See the preceding note.

[10] One might think, of course, that explaining away one kind of irrationality by making it into another is not such a promising move. Self-deception after all might not seem any less mysterious than weakness of the will. I am not sure about the respective degrees of mysteriousness, but it seems easier to accommodate one kind of irrationality instead of two. A similar strategy is employed by Mele in Mele 1987. He takes it in the other direction, however; he tries to understand self-deception as a kind of doxastic weakness of the will.

[11] Another relevant passage is the following: "If *r* is someone's reason for holding that *p*, then his holding that *r* must be, I think, a cause of his holding that *p*. But, and this is what is crucial here, his holding that *r* may cause his holding that *p* without *r* being his reason; indeed, the agent may even *think* that *r* is a reason to reject *p*" (Davidson 1970, 41, emph. mine). 'Being someone's reason' here seems to be construed as *being thought of as a reason by that person*, that is, construed as requiring second order belief. This passage is a bit puzzling, too, for it is left unclear whether *r* is *a* reason for holding that *p* at all. To be relevant for akrasia, however, it has to be. So, assume that it is. Then, it would follow that *r* is *the* reason for which the person holds that *p*. (It would follow, that is, given that it causes holding that *p in the right way*. But causing holding that *p* despite a second order belief that *r* is not a reason for holding that *p* is, I think, not excluded by the in-the-right-way clause; that clause is supposed to deal with internal deviant causal chains.) Nevertheless, it might not be the reason

Davidson rescues principles P1 and P2 from akrasia by the distinction between prima facie and all out judgments. This means that they hold for all-out judgments or intentions only. The connection between intention and action is kept tight, so to speak, in order to save the principles, while the connection between reasons and intentions is losened. Davidson is, of course, fully aware of this:

> Intentional action, I have argued in defending P1 and P2, is geared directly to unconditional judgments like 'It would be better to do a than to do b'. Reasoning that stops at conditional judgments (...) is practical only in its subject, not in its issue (Davidson 1970, 39).

Now consider: If I judge that *x* is better than *y*, that means that I want *x* more than *y*. My action *x*, that is, shows that the value I set on *x* was higher than that of *y*. This, too, is a consequence Davidson explicitly endorses; he says about akratic agents: "their intentional action shows that they have set a higher value on the act they perform than their principles and their reasons say they should" (Davidson 1985a, 140). In the akratic case, that is, I set a higher value on the action than results from my practical reasoning. But from my practical reasoning, it results that, given all my reasons, *y* is better than *x*. And to this judgment, P2 clearly does not apply. The result of my practical reasoning, that is, does not necessarily correspond to my real wants.

Now, in a case of pure akrasia, we can assume that the agent does not commit any mistake in his reasoning. Given his beliefs, desires, and the values he attaches to his desires, that is, he 'calculates' correctly. On this assumption, the discrepancy between the outcome and his real desires must originate in a corresponding discrepancy further up, in a discrepancy, that is, in the premises of the practical reasoning. Consequently, on this account, no prima facie judgment necessarily corresponds to our real desires.

But then, what does such a judgment express? In the passage just quoted, Davidson talks about the value I "should" set on an action given my principles and reasons. But I don't think that makes tremendously much sense for a Davidsonian. On the contrary, it means importing a normative-explanatory ambiguity into an otherwise admirably pure explanatory concept of evaluation. According to Davidson, an akratic action violates the result of internal, topic-neutral practical reasoning, not of practical reasoning that is read normatively in

---

for which he *believes* he holds that *p*. In this case, the second order belief is the *only* irrational state involved, for given his first order reasons, the person's holding that *p* is perfectly rational. Something similar seems to me to hold for the self-deceived akratic agent: what he does is, given his real desires and beliefs, perfectly rational. What is irrational is only what he believes about his desires.

any of the ways described above. Thus, the value I "should" set on an action *given my reasons* simply *is* the value that correctly results from 'crunching' those reasons according to the Davidsonian calculus. This is supposed to simply be a function of my actual beliefs and desires. Not of my actual beliefs and those desires that I *should* have.

But if I was not calculating in terms of any kind of thicker evaluation, I must have made *a mistake* somewhere. Since, by assumption, the mistake is not in the reasoning itself, in the inferences drawn, it must be in the very premises. It must be a mistake as to *how much I actually value* something. Even on Davidson's own terms, that is, akrasia turns into self-deception. His own account, if stared at long and hard enough, turns out to be a version of the very kind of skepticism about weakness of the will sketched at the beginning of this paper.

By way of concluding, I would like to point out two further consequences of this result. First, on the face of it, it might seem possible enough to make mistakes about the strength of our own desires. Especially, maybe, in predicting them. On second sight, however, this result points to another tension in this picture of practical reasoning. Processes of practical reasoning do not have to be conscious, according to Davidson; it's our beliefs and desires, consciously considered or not, that determine our intentions. Practical reasoning according to the calculus described is supposed to model both kinds of decision making: those that involve conscious deliberation and those that do not. In cases without conscious deliberation, what figures in this process are the very first order states themselves, however. That is, it is the desires themselves that, together with our beliefs, condition the resulting value. And a desire cannot be mistaken about its own strength, it simply has a certain strength. The kind of self-deception required for actions that violate better judgments, that is, can occur only where the practical reasoning in question takes the form of conscious deliberation.

In these cases, however, the action actually performed is not irrational in any practical sense. It does not violate the reasons the agent actually has. Therefore, akrasia, secondly, does not force us to keep the logical gap between reasons and actions open in the way Davidson does. All we need is the possibility of mistaken beliefs about (the strength of) our desires. Such belief is mistaken exactly because the desire in question is involved in the causation of action according to its real strength. Therefore, what he actually does will show what an agent valued most. There is no need to losen this connection.[12]

[12] Without the logical gap, a perennial worry about the explanatory force of (Davidsonian) reasons-explanation could maybe be put to rest. For the gap allowed for having reasons of the kind supposed to explain actions of a particular kind without performing an action of that kind. Thus, reasons-explanations appeared to be, at best, incomplete explanations. Even the

Moreover, realizing that akrasia on this model really is a kind of self-deception shows what was apt in the feeling voiced by many commentators that Davidson allows akrasia in only one place, while it actually can break practical rationality almost anywhere. Particularly, it has been felt, akrasia violating one's own intentions should be possible.[13] Now, if akrasia really is self-deception about desires, such self-deception would indeed seem equally possible for any kind of desire — regardless of how tight the connection between the real desire and the action is. And according to the Davidsonian model, intention is a kind of pro-attitude or desire.

Let's sum up: Weakness of the will threatens an account of action explanation insofar as it seems to cut the very connection between reasons and action that is constitutive of intentional action. The better judgment in terms of which our tradition defines akrasia accordingly employs a purely internal, topic-neutral concept of evaluation. For only reasoning in terms of such a concept is explanatory in the sense required for intentional action. Getting clear about the nature of this concept of evaluation, however, at the same time shows that the conflict is spurious; actions of the required kind are impossible after all. This result does not only apply to the Davidsonian account of weakness of the will. Rather, it holds for any account that requires intentional actions to have reasons-explanations. Such skepticism about weakness of the will, defined as violation of an internal, topic-neutral better judgment, does not commit us to denying that there are, in fact, cases of irrationality in which it seems even to the agent himself that he acts against such a better judgment. Even though many seemingly paradigmatic cases might turn out to be misdescriptions owing to the concept of evaluation employed being a thick, normative one, others might withstand such diagnosis. The particular irrationality involved here, however, is not of the practical kind defined as akrasia. Rather, these are cases of a particular kind of self-deception: self-deception about desires and their strengths.[*]

.

---

modified model of practical reasoning was not able to determine the causally effective reason. Cf., for instance, Mele 1987, 89. On a purely explanatory reading of Davidsonian practical reasoning, however, the calculus of practical reasoning would determine effective reasons; it would describe what agents do given that they have certain reasons. This might ultimately undermine the motivation for the claim that reasons-explanations are causal explanations. For, after all, this was motivated by an inability to determine effective reasons. But that is a topic for another paper.

[13] Cf. Audi 1979, Pears 1984, Walker 1989.

[*] I would like to thank Peter Pagin, Lilli Alanen, Fred Stoutland, Jonas Olson and listeners in Uppsala and Umeå for helpful comments.

## *References*


Audi, R., 1979, Weakness of Will and Practical Judgment, *Nous* 13, 173-196.

Davidson, D., 1970, How is Weakness of the Will Possible?, reprinted in Davidson 1980, 21-42.

—, 1978, Intending, reprinted in Davidson 1980, 83-102.

—, 1980, *Essays on Actions and Events*, Oxford: Oxford University Press.

—, 1982, Paradoxes of Irrationality, in *Philosophical Essays on Freud*, ed. R. Wollheim, J. Hopkins, Cambridge: Cambridge University Press, 289-305.

—, 1985, Incoherence and Irrationality, *Dialectica* 39, 345-354.

—, 1985a, Deception and Division, in *Actions and Events. Perspectives on the Philosophy of Donald Davidson*, ed. E. LePore, B. McLaughlin, Oxford: Oxford University Press, 138-148.

—, 1985b, Replies to Essays I-IX, in *Essays on Davidson. Actions and Events*, ed. B. Vermazen, M. B. Hintikka, Oxford: Clarendon Press 1985, 195-229.

Frankfurt, H., 1971, Freedom of the Will and the Concept of a Person, *Journal of Philosophy*68, 5-20.

Jackson, F., 1984, Weakness of Will, *Mind* 93, 1-18.

Jeffrey, R., 1974, Preference among Preferences, *Journal of Philosophy* 71, 377-391.

Lazar, A., 1999, Akrasia and the Principle of Continence or What the Tortoise Would Say to Achilles, in *The Philosophy of Donald Davidson (The Library of Living Philosophers Vol. XXVII),* ed. L.E. Hahn, Chicago and LaSalle, Ill: Open Court, 381-401.

Mele, A., 1987, *Irrationality*, Oxford: Oxford University Press.

—, 1992, Akrasia, Self-Control, and Second-Order Desires, *Nous* 26, 281-302.

Peacocke, C., 1985, Intention and Akrasia, in *Actions and Events. Perspectives on the Philosophy of Donald Davidson*, ed. E. LePore, B. McLaughlin, Oxford: Oxford University Press, 51-74.

Pears, D., 1984, *Motivated Irrationality*, Oxford: Clarendon Press.

Rorty, A., 1983, Akratic Believers, *American Philosophical Quarterly* 20, 175-183.

Schiffer, S., 1976, A Paradox about Desire, *American Philosophical Quarterly* 13, 195-203.

Walker, A. F., 1989, The Problem of Weakness of Will, *Nous* 23, 653-676.

Watson, G., 1977, Skepticism about Weakness of Will, *The Philosophical Review* 86, 316-339.