INVERSE AKRASIA AND WEAKNESS OF WILL

RICHARD HOLTON, MIT

Abstract: The standard account of weakness of will identifies it with *akrasia,* that is, with action against one's best judgment. Elsewhere I have argued that weakness of will is better understood as over-readily giving up on one's resolutions.

Many cases of weak willed action will not be akratic: in over-readily abandoning a resolution an agent may well do something that they judge at the time to be best. Indeed, in so far as temptation typically gives rise to judgment shift—to a tendency to change one's judgment so that one values the tempting option as the best—weak willed action will typically be akratic. But conversely, strong willed action now looks as though it will be akratic. I argue though that it need not be, once we distinguish between actual judgment, and dispositions to judge.

Within this framework, the issue of 'inverse *akrasia*' looks rather different. I argue that whilst Huckleberry Finn plausibly does show weakness of will in abandoning his resolve to turn Jim in, it is far from clear that he is akratic. Whilst cases of inverse *akrasia* are clearly theoretically possible, I suggest that, given cognitive dissonance mechanisms, they are unlikely to be very common.

Can akratic action—that is, action against one's best judgment—be rational? It looks as though it must involve at least *some* irrationality. Think what is at stake. We suppose the agent judges that a certain action is best—not just morally best, or prudentially best, or best by the standards of the society in which they live—but best overall, in the light of all the considerations that they can bring to bear upon it. We then suppose that they knowingly, intentionally, voluntarily act contrary to that judgment. So how could they fail to manifest a degree of practical irrationality? Consider the theoretical analogue: the agent judges that they should form a certain belief, and then they form another belief, inconsistent with it. What clearer case of theoretical irrationality could there be?

Of course in both the practical and the theoretical cases the irrationality might be off-set in some way. It may be that there is good reason for acting as they do, reason that their judgment has somehow failed to register; if this is so their action may manifest a degree of rationality too. It may be that the action they perform is actually the best, or, in the theoretical case, that the belief they form is true. These are the cases that involve what Arpaly and Schroeder have dubbed 'inverse *akrasia*'.[1]

[1] (Arpaly and Schroder, 1998)

But to say that they show some rationality in these cases is not to say that they show no irrationality. All we can conclude is that we can get ourselves into situations from which there is no fully rational way out.

Doubtless there is more that should be said about this issue, but I shall not say it here. Here I accept that truly akratic action is, to some degree at least, irrational. My aim is to investigate how worrying this should be. my contention is that it is less of a problem than might be supposed, since truly akratic action is far less common than the large philosophical discussion might lead one to believe. For a start, there is good evidence that temptation typically give rise to *judgment shift*: what would have been judged the less good action is judged the better, once the time comes to act. As a result, action that would have been akratic, now comes to be in conformity with best judgment. But of course, this merely flips the problem, for now it will seem that it is the person who maintains a resolution in the face of temptation who will be acting akratically: in maintaining the resolution they will not be doing what they now judge best. In answer to this I want to stress how slippery the notion of judgment is. In the examples they give philosophers frequently suggest that there is a simple fact of the matter about what people judge best. I doubt that that is often so. It is in the nature of the putative cases of akratic action that judgment is conflicting, unclear, and unstable.

First, we need to distinguish between explicit judgments, and dispositions to judge, since both go under the heading of 'belief'. Judgment shift is often purely dispositional. The evidence suggests that those who are successful in maintaining resolutions do so by not reconsidering them. Dispositional judgment shift is never made actual. This mean that in judging the rationality of what is done, we need to assess the rationality of non-reconsideration. And this is something that should be assessed by pragmatic standards, so that it can be rational not to reconsider even if, were one to reconsider, it would be rational to change one's mind.

Second, even in cases in which judgments are explicitly made, their content can be far from clear. They may be half-hearted, or equivocal, or contradictory. In developing this thought I will reexamine the case that has figured so centrally in discussions of inverse *akrasia*: that of Huckleberry Finn, and of his loyalty to the fugitive slave Jim, despite the protests of his conscience that he should hand him in.[2] The case is such a fruitful one largely because, as Jonathan Bennett says in the discussion that brought it to philosophical prominence, Twain wrote it 'with such unerring precision'.[3] My contention is that when one looks with equal precision at what Twain wrote, one should not conclude that it is a clear case of *akrasia* at all. There is a point at which Huck believes that he should turn Jim in, but when once he starts to act on that belief, his conviction is rapidly lost. What remains is the shouting of his conscience; but he certainly does not take this as the last word on

[2] (Twain, 1884). The passages most cited are in Chapter XVI, but the issue comes up repeatedly throughout the book.

[3] (Bennett, 1974)

the subject. This comes out all the more clearly when we look at some additions to the relevant chapter of Huckleberry Finn that Twain made for a later reading tour. But what is brought out is, I think, quite clearly in the original.

Before I do any of this though, I need to set out the framework within which I shall be working, starting with how I understand the relation between *akrasia* and weakness of will.

AKRASIA AND WEAKNESS OF WILL

Standardly *akrasia*—action against one's best judgment—has been identified with weakness of will. Elsewhere I have argued that these are really two distinct notions, though they often overlap in extension. Weakness of will, I have suggested, involves an over-ready revision of one's resolution, where a resolution is a special sort of intention—an intention that is designed to stand firm against later contrary inclinations.[4] Such inclinations might come from different sources: they might come from temptation; or from fear; or even, as in the case of Huck Finn, from the prompting of one's conscience. Anticipating such inclinations, agents form intentions that are designed to work against them.

Not every revision to a resolution is an over-ready revision. Sometimes we are right to reconsider our resolutions: our circumstances or our projects might change, or we might come across new information. And sometimes, when we reconsider, we are right to revise. To deny this is not to be strong willed, but simply to be stubborn.

So strength of will involves walking a fine line between weakness, on the one side, and stubbornness on the other. How is this to be achieved? We might expect that it comes from having a good grip on the reasons for the resolution, a grip that enables us, when we reconsider, to reaffirm the resolution if things have not changed significantly. But a large body of empirical work indicates that successful resolution is rarely like that. In a series of experiments, Walter Mischel and his colleagues investigated the factors that enable children to resist an immediate gratification in order to get something better later.[5] They expected to find that being reminded of the rewards for holding out would serve to increase the children's resolve. But they found that, on the contrary, it great increased the likelihood that they would succumb to the temptation of the immediate reward. Successful resistance does not typically stem from successful reaffirmation; rather it comes from resisting

---

[4] (Holton, 1999), (Holton, 2003). Recently Al Mele has conducted some studies that he takes to show that what American subjects mean by 'weakness of will' is actually closer to what I mean by '*akrasia*' (Mele, forthcoming). My British ear remains sceptical, but even if he is right, the central point is that there are two notions here, no matter how they are picked out in standard English.

[5] For summary see (Mischel, 1996).

reconsideration in the first place. Conversely, succumbing to temptation typically results from reconsideration. How is this to be explained?

JUDGMENT SHIFT[6]

Building on Mischel's work, Rachel Karniol and Dale Miller explored what was happening to the children's judgments during such experiments.[7] Eight year-olds were shown marshmallows and chewing gum and asked which they preferred. Half the children were then told that they could have their first choice, but only after the experimenter returned from some tasks she had to do; these provided the controls. The other half were told the same, but were told in addition that at any point they could ring a bell to summon the experimenter, in which case they would get their second choice. The marshmallows and chewing gum were left in plain sight.

After ten minutes the experimenter returned (the few children who rang the bell in the meanwhile having been excluded). She told the children that she was not yet in a position to give them their rewards, but that she needed to ask a further question, one that she forgot to ask before. The question concerned the value, on a scale of one to five, that they placed on the two options. And here is the interesting finding: the group who had the chance to ring the bell gave a value to their preferred choice that was significantly lower than that given by those who did not have the chance to ring the bell.[8] Moreover, it was the group who had the chance to ring the bell whose valuations were anomalous. The valuations of those who had no chance to ring the bell are the same as those of a third control group who did not have to wait; and the same as those of fourth and fifth groups, who were treated just like the first and second group respectively, with one difference: the marshmallows and chewing gum were not left in their sight.

So what is happening? It seems that children who (i) had the chance to ring the bell, and (ii) kept the options in sight, came to devalue the option that was initially preferred. But that is just the case in which there is maximum temptation: the children had the possibility of giving up a later, greater benefit for a lesser, more immediate one; and that possibility was made very salient by the visual presence of the options. Moreover, it really does seem to be the temptation that is doing the work. In a subsequent experiment Karniol and Miller found that the phenomenon did not occur when there is a large difference in the initial valuation: if something is not attractive enough to be a real temptation, it does not lower the value placed on the thing it is competing with.

This is a case of what I shall call *judgment shift*. The temptation causes the agent to reevaluate: the value of what will be gained by holding out goes down, and so the

[6] This section draws on Chapter 5 of (Holton, 2009).
[7] (Karniol & Miller, 1983).
[8] There is no difference in the value of their less preferred choice.

relative value of what will be gained by succumbing goes up.[9] By the time agents succumb their action will typically not be akratic, since their judgment about what it is best to do will have followed their judgment of which outcome is most desirable. Of course this will seem a trivial example (though perhaps not to eight year-olds). But there is nothing special about the experimental set-up; we have good reason to think that this is a very general phenomenon. When we succumb to temptation we tend to judge that that is the best thing to do.

The most obvious explanation of what is going on comes from cognitive dissonance theory. In general we work very hard to ensure that the picture we have of ourselves is coherent: that it is not 'dissonant'. Moreover we want—are *driven*—to come up with a picture that puts us in a good light.[10] Achieving this involves us in all kinds of change of attitude: among other things, people can reinterpret how interesting something is, how much it hurts, how good it is, and how much they want it; all of this can change if the change makes for a more coherent, flattering picture. And, importantly for the case at hand, reinterpreting a state frequently leads to a real change in the state being reinterpreted. People actually come to want something less—they will sacrifice less to get it—because they interpret themselves as valuing it less.

So here is the proposed understanding of the Karniol and Miller finding. The children who undergo judgment shift are tempted to ring the bell to get the less desirable option, rather than the waiting for the more desirable option. They are aware of the temptation growing, and aware that they are likely to act on it. So they start to tell themselves a story that will make sense of this behaviour. One story would be of the kind that economists tell, one involving, implicitly at least, discount rates. They could come to think that they prefer the lesser good now, rather than the greater good later, because they add in the unpleasantness of waiting. But that is a complicated story, and, besides, it leaves them open to later regret. For they will later think that if only they had not had such a steep discount curve—had not been so impatient, in ordinary talk—they would have got something they liked more. Better to simply reevaluate the two options themselves, independently of the wait, so that the one available is the one that is preferred. Then one will be able to have now

---

[9] But not the absolute value, which remains significantly unchanged; see (Karniol & Miller, 1983) p. 938.

[10] Establishing how the desire for coherence and the desire to look good interact—in particular, assessing which one is dominant in those cases in which they conflict—is controversial. It is further complicated by the question of what the standards are against which individuals want to look good: their own, or those of the group around them. For a useful history of the cognitive dissonance approach see (Cooper, 2007). Following his own 'Self-Standard Model', Cooper argues that a desire to look good by community standards is particularly central. For details see (Stone & Cooper, 2001).

what one most wants, and there will be no later regret. This, I suggest, is what is happening here.[11]

Assuming that this is right, then the change in valuation is not the *origin* of the process that leads to the subjects yielding to temptation: it is rather itself caused by the children's awareness that they are likely to yield. Nevertheless the change in valuation is real. Indeed, although Karniol and Miller did not do the experiment, there is some circumstantial evidence for thinking that, at the point at which the children would succumb to taking what they formerly viewed as the second best option, they would actually choose it over the other option if offered a free choice between the two (though it is also possible that a free choice would lead to a reconsideration that would restore the original valuation). My reason for saying this it that it has been shown (i) that reducing subjects' resolve, so that they take a tempting but otherwise less preferred option, leads them to choose that option even for circumstances in which their resolve would not be reduced[12]; and (ii) that once a choice has been made, subjects come to strongly prefer the thing they have chosen, even if at the time of the choice the preferences were very close.[13]

If the change in valuation is not the source of the process that leads to yielding, what is? What causes the subjects to yield is desire, in one sense of that rather broad term. It is the desire for the sweet that is available now. We can get some purchase on its nature by recalling that the change in valuation does not occur when the sweets are covered—when, out of sight, they can be kept out of mind. In contrast, when something is visually present it is much more likely to come incessantly into one's thoughts, and this is an important factor in desire. Incessant presence, of course, is not enough to constitute desire. Thoughts of a dreaded thing will occupy one's mind. Nevertheless, it is central. We get closer to characterizing the state we are after if we turn to Scanlon's notion of *desire in the directed-attention sense*: this happens, he says, when the thought of an object 'keeps occurring to him or her in a favorable light, that is to say, if the person's attention is directed insistently toward considerations that present themselves as counting in favor'.[14]

Yet that still is not quite right. In a guilty state of mind, the things that count in favour of the virtuous but forsaken course of action may come insistently to my attention; but that does not mean that I want to take it. What is missing in Scanlon's characterization is the idea that desire *pulls* me to a course of action: that I have an *urge*, or, in more extreme cases, a *craving*, something that moves me to do it. Such a feature cannot, I think, be reduced to more cognitive talk of focusing on an

---

[11] Oddly Karniol and Miller seem to give the first interpretation (p.936); but the children are evaluating the things themselves, not the packages of the things *together* with the wait. Of course, it could be that they are unconsciously factoring in the wait; but we would need evidence that this was so.

[12] (Wang et al, forthcoming).

[13] This was one of the core early findings of the cognitive dissonance approach (Brehm, 1956).

[14] (Scanlon, 1998) p. 39.

object or seeing things in a certain light. These cognitive features may be necessary for desire to arise; but they do not constitute it. Desire in the sense we are after is a state that preoccupies an agent's attention with an urge to perform a certain action.

So, to sum up, what I think is happening in the Karniol and Miller experiments is this: the tempted children find their attention focused on the immediately available sweet; as a result they find themselves with a strong urge to ring the bell to get it; and, as they become aware that they are likely to succumb to this urge, they change the evaluation of their options so as to avoid cognitive dissonance.

I do not want to say that this happens in every case of temptation. And I am certainly not making the (implausible) analytic claim that whatever an agent chooses simply is what they value most highly. Rather I am describing a causal process. So there are surely cases of ordinary temptation (i.e. cases not involving addiction) in which this process does not take place: cases in which agents choose an option whilst, even at the time, valuing that option less than some other that is available. I suspect though that the power of the cognitive dissonance mechanism is so great that such cases are unusual. Even where agents act in ways that they initially know to be morally wrong, the ability to reinterpret what is happening to put themselves in a reasonable light is remarkable.[15] As a result, cases of *akrasia* are rarer than most philosophers have supposed.

THE RATIONALITY OF MAINTAINING RESOLUTIONS[16]

If judgment shift is a pervasive phenomenon, it puts the rationality of weak willed and purportedly akratic actions in a rather different light. For standardly when they occur, those actions will no longer be akratic. They may have been anticipated as akratic, but by the time they happen the judgment will have shifted so that they will be in conformity with the agent's best judgment. I suspect that something along these lines happens in purported cases of reverse *akrasia* too. But before addressing that, I want to consider what happens in cases where agents manage to maintain their resolutions. For if succumbing involves acting in accordance with one's current best judgment, doesn't holding out involve acting against it? Let us call this *the problem of akratic resolution*.

My approach to the problem involves embracing a two-tier account of resolutions, along the lines of that advocated by Michael Bratman for the case of intentions in general.[17] Let us follow Bratman's reasoning there. His central idea is that it can be rational to have a general policy of not reconsidering intentions in

[15] For discussion of a large range of real cases see (Baumeister, 1996) Ch. 2. Perpetrators, even of the most horrific crimes, typically see themselves as victims.

[16] This section draws on (Holton, 2004)

[17] Note though that Bratman does not endorse this approach for the special case of resolutions; instead he invokes a further constraint in terms of lack of regret. I argue against this in (Holton, 2004)

certain circumstances. This policy can confer rationality on one's action when one acts on a particular intention, rationality that that action might not otherwise have. In order to confer this rationality, Bratman argues, it must have been rational to form the intention in the first place, and it must have been rational not to revise it at each point between its formation and the time of action.[18]

The thought here isn't that forming an intention gives an extra reason to follow through with that intention. However, whilst intentions don't create new reasons for the action, they do entrench the decisions that are arrived at on the initial consideration, since they give *reasons for not reconsidering*. If the agent had not earlier considered what to do, they would now have reason to consider; but their earlier consideration provides a reason for not considering again.

The entrenchment that intentions provide is defeasible: sometimes things will change so radically from what was expected that it will be rational to reconsider the intention. However, provided things do not change radically, it will be rational to go ahead with the intention without reconsidering. This gives the possibility of Nietzsche's "occasional will to stupidity": sometimes one will follow courses of action that would seem stupid if one were to have reconsidered.[19] But by and large not reconsidering is beneficial. It enables economy of effort (I consider once, and then do not waste scarce time and effort in further consideration); and it provides coordination advantages (having fixed an intention, my other actions, and the actions of others, can be coordinated around it).

It might be thought that to embrace the two-tier strategy is to accept that it is rational to make oneself irrational. That, I think, is a mistake. I would be irrational if I reconsidered an intention, and decided to stick with it even though the reasons I then had went against it. But the whole point is that there is no reconsideration; to reconsider would defeat the point of having intentions. Indeed, very often I do not even consider whether to reconsider. I simply have unreflective habits that determine when to reconsider, and when not.

A more plausible line of objection is that the two-tier strategy makes our actions *arational*: since we do not reconsider, rational assessment simply does not come into it. Certainly there are ways of sticking with intentions that do involve making oneself arational. If I intend to stay in the same place for the next six hours, a powerful sleeping drug will do the job at the price of making me arational for that period. However, that is not the model that we are proposing. There are good reasons for thinking that agents who employ a strategy of nonreflective nonreconsideration do not thereby make themselves arational. First, rationality concerns what we have the *capacity* to do. In employing a habit of nonreflective nonreconsideration we do not make ourselves unable to reconsider. We still *could* open the question up again, even if circumstances do not change. It is just that we

[18] (Bratman, 1987) p. 80.
[19] (Nietzsche, 1886) §107.

do not. (In developing the skill of catching a ball I do not make myself unable to drop it.) Second, employing a habit of non-reconsideration does not involve completely closing down one's faculties. We still engage in lower level thought about how the intention is to be implemented; and we still need to monitor to ensure that things have not changed so radically that the intention requires reconsideration after all. Although this monitoring will often be non-reflective, it is still a rational process.

Can we apply the two-tier account to resolutions? My main contention here is that we can. The idea, of course, is that resolute agents acquire the disposition not to reconsider resolutions, even though, were they to reconsider, they would revise them. In many cases such revisions would be rational, by the lights of the agent at the time: their judgment about what it would be best to do would have changed. Yet despite this potential judgment shift, the failure to revise would not be irrational since it would result from a policy of non-reconsideration that was itself rationally justified on pragmatic grounds. The earlier consideration, and the resolution that came from it, provide a reason for not now reconsidering.

Again it might be objected that, in training oneself not to reconsider resolutions, one makes oneself arational. The issues here are exactly parallel to those for intentions in general. Certainly there are strategies for resisting temptation that involve making oneself arational; again, sleeping through the temptation is one.[20] But having the disposition not to reconsider resolutions need not be among them. It need not involve losing the capacity to reconsider; indeed, keeping oneself from reconsidering will often involve effort. Furthermore, pursing a policy of non-reconsideration doesn't involve switching one's mental faculties off. Normal intentions come with thresholds beyond which reconsideration will take place. Certainly for resolutions any such thresholds should be set very high: otherwise the corrupting effects of temptation on judgment will make the resolutions all too easily broken. Nevertheless, some such thresholds are surely needed; there is no point in persisting with one's resolution to exercise if one discovers that exercise is actually damaging one's health.[21] Equally importantly, we need to survey our resolutions to ensure that they are being implemented. This is especially so where we are trying to overcome habits—like smoking or sleeping in—that are so deeply ingrained that the actions become automatic.[22]

---

[20] This is the strategy used by one of the children in Mischel's delayed gratification experiments described above. See (Mischel, 1996) p. 202.

[21] We might here distinguish pressure for revision coming from the very inclinations that the resolutions were designed to overcome, from pressure coming from other sources: genuinely new information, for instance. Perhaps the thresholds should be sensitive only to the latter sort of pressure.

[22] For discussion of the importance of such self-monitoring see (Carver & Scheier, 1998).

THE PROBLEM OF AKRATIC RESOLUTION

We can now address the problem of akratic resolution directly. The problem, recall, is that in cases of judgment shift it seems that to act resolutely will be to act akratically; and that appears irrational.

The problem of akratic resolution is an instance of a general problem about whether it can be rational to be akratic. There is little doubt that acting akratically can sometimes be the most rational course of those available: the judgments against which one acts might be crazy. The question is rather whether it nonetheless necessarily involves a degree of irrationality. Recently a number of authors have argued that it need not. To take one example: it is clear that our emotional responses can track reasons that we fail to notice in our judgments; and hence some have concluded that it can be rational to be moved by these emotions even when they run contrary to our judgments. We might, for instance, have an emotional sense that we should not trust a person, and this sense might be reliable, even though our explicit judgment is that the person is quite trustworthy. [23]

As I mentioned at the outset, this might be right; but it is far from obvious that it is. It certainly seems as though if one makes a serious and considered judgment that a certain action is, all things considered, the best, it will involve a degree of practical irrationality to act against that.[24] It seems that this is the practical analogue of believing something when one thinks the evidence is against it; and that seems to involve irrationality, even if one's belief is true. In the case of vendettas it can be beneficial to be irrational.[25] Why isn't this just another instance of the same thing? At most is seems that we have distinguished a new sense of rationality: an externalist, reliabilist sense, in which acting against one's best judgment is not irrational, to set against the internalist sense in which it is.

I cannot resolve the general issue between internalist and externalist conceptions of rationality here. What is important for us is that the two-tier account simply sidesteps the problem. For if agents do not reconsider, they do not ever form the judgment against which their resolution requires them to act. In the face of temptation they have the disposition to form those judgments, but the disposition is not realized. The judgment shift is merely potential. So they are not akratic.

In saying that agents do not reconsider, I do not mean that they do not think about the issue at all; as we have seen, some thought will typically be necessary for effective monitoring. Non-reconsideration only requires that they do not seriously reopen the issue of what to do, and seriously arrive at a new judgment. Nonetheless, it might seem that this makes rationality far too fragile. I am arguing that

---

[23] (McIntyre, 1990); (Anthony, 1993); (Anthony, 2000); (Arpaly, 2000). For a criticism of some features of the approach of these writers (though not of the overall conclusion) see (Jones, 2003).

[24] For a presentation of the internal ('narrow') conception of irrationality, see (Scanlon, 1998) pp. 25 ff.

[25] (Schelling, 1960)

rationality can be preserved provided that the agent does not form the all things considered judgment that it would be best to abandon the resolution. Yet mightn't the agent form that judgment without reconsidering what to do? A little too much thought in the wrong direction, and the agent will fall over the abyss into irrationality. This in turn will mean that irrationality will be very frequent. For surely it is part of the nature of temptation that judgment shift is frequently not merely potential, but actual.

But this is to misunderstand the nature of temptation. It is certainly true that, prior to any reconsideration, temptation brings new, or newly strengthened, desires. It is also true that it will bring new judgments: the judgments, for instance that abandoning the resolution will not have some of the bad consequences previously envisaged, or that it will bring unforeseen benefits. Yet such judgments fall far short of the judgment that it would be best, all things considered, to abandon the resolution. That judgment involves not just an evaluative judgment, but a comparison: a *ranking* of one option as better than the others. And that ranking is not an abstract, impersonal one; it is ranking of options as options for the agent. Such a ranking is not easily arrived at. It requires real mental activity from the agent. It is not the kind of thing that simply arrives unbidden. [26]

I think that this is enough to rebut the fragility worry. But I want to go further, and suggest that there is an even stronger reason for thinking that we will not arrive at new all things considered judgments in the absence of reconsideration of what to do. How do we form all things considered judgments? I suggest that, standardly, we form them by deciding what to do.[27] That is, rather than thinking that we first arrive at a judgment about what is best, and then decide what to do on the basis of that judgment, things are the other way around. We start by deciding what to do, and then form our judgment of what is best on the basis of that decision. This is not to say that the judgment about what is best is identical to the decision about what to do; we know that we might have made a mistake in our decision so that it does not correspond to what is best, a possibility made all the more vivid by reflecting on our own past decisions, or those of others. It is simply that one's best way of deciding which action is best is via serious consideration about what to do.[28]

---

[26] I speak of judgments, rather than of beliefs, because of a strong tendency in philosophy to think of beliefs dispositionally: what one believes is what one would judge if one were to consider the matter. But that is exactly to obscure what is at issue here. These are cases in which agents would arrive at different judgments if they were to consider the matter at different times; and the question is whether they should go in for such consideration. I suspect that, in a desire to avoid a certain crude reified picture of both beliefs and desires, philosophers have in general moved too far towards dispositional accounts. Our dispositions are simply not stable enough to support beliefs and desires understood in this way: they are far too sensitive to framing effects.

[27] I discuss this idea in more detail in (Holton, 2006).

[28] There is a parallel here with the much discussed phenomenon that one's best way of determining whether one believes that p is simply by doing one's best to determine whether or

I do not claim that it is impossible to reach a judgment about what is best except via a judgment about what to do. In psychology few things are impossible. There are, for instance, reckless agents who know that their decisions about what to do are no guide to what is best; and there are depressed agents whose will is paralyzed, so that they judge what is best without being able to bring themselves to decide to do it. It is enough for my purposes if the typical, nonpathological, route to best judgment is via decision about what to do. For that will guarantee that, in the typical case, the only route to a new judgment about what is best is via a reconsideration of what to do. So if agents do not reconsider, they will not arrive at new judgments, and will not be akratic. Rationality is even less fragile than was feared.

What of those cases in which the agent does arrive at the judgment that it would be best to succumb? This might happen, unusually, without the agent reconsidering what to do: perhaps the immediate judgment shift is so enormous that the agent can see no benefit whatsoever in persisting with the resolution (I take that such cases are very unusual: whilst temptation often leads us to believe in the advantages of succumbing we normally retain a belief that there is *something* to be said for holding out). Alternatively the agent will reconsider what to do, and will make a judgment that it is best to succumb as a result of that reconsideration. In such circumstances, would persisting in the resolution involve irrationality? Addressing this takes us straight back to the general problem of the irrationality of *akrasia*. I suspect that it will: that even if persisting in the resolution is the most rational course, some local irrationality will be required if they are to get themselves out of the problem into which their revised judgment of what is best has led them.

The two–tier account thus does not ascribe rationality in every case; but it does provide a promising explanation of how maintaining a resolution will typically be rational. It is particularly attractive since it chimes so well with the empirical work on how we in fact stick by our resolutions: the primary mechanism, as we have seen, is exactly that of avoiding reconsideration. Even thinking about the benefits to be gained by remaining resolute makes an agent more likely to succumb. Once we have resolved, the best plan is to put things as far out of mind as possible.

BACK TO HUCKLEBERRY FINN

So weakness of will need not be akratic; nor need strength of will. What consequence does this have for the phenomenon of inverse *akrasia*? I want to suggest

---

not it is the case that p. Here again, although one provides a route to the other, we recognize that the two states are different, since one's beliefs can be false. See (Moran, 2002) pp. 60 ff. for a nice discussion. The parallel, however, can be taken too far: in some sense the belief case is the opposite to the case of practical deliberation. In the former one looks to the world to discover a truth about oneself; in the latter one looks to oneself to discover a truth about the world.

that the standard cases are not really cases of akrasia at all. Rather they are cases of non-akratic weakness of will. Or, at least, they do not clearly involve *akrasia*. They involve judgment shift, but not motivated by simple desire as is normal cases of temptation. The influence of desire might be there, but so too is the attempt, again underpinned by cognitive dissonance mechanisms, to make sense of how one acts.

It is hard to give a generally argument for this, so instead I shall say something anecdotal, returning to the passages of Huckleberry Finn in which Huck wrestles with the issue of whether to turn Jim in.[29] Bennett sees it as involving a clash between judgment and sympathy: Huck's judgment that he should turn Jim in, fighting with his sympathy for Jim. Since it is his sympathy that wins out, it is for Bennett a clear example of *akrasia*. Many other philosophical commentators have concurred.[30]

However, this was not quite how Twain himself saw the case. In a series of performances across North America, Australia and India, Twain read the episode from Chapter xvi, wrote an introduction to it, and amended the text, adding a number of lines to bring out the dramatic force.[31] 'In a crucial moral emergency' he wrote in his introduction, 'a sound heart is a safer guide than an ill-trained conscience.' That doesn't sound as though the clash is between moral judgment and sympathy. It sounds rather as though there is more that one way of getting at the moral truth, and that an ill-trained conscience does it rather badly. 'It shows' he went on, 'that that strange thing, the conscience—that unerring monitor—can be trained to approve any wild thing you want it to approve if you begin its education early & stick to it.'

Indeed, Twain thought rather badly of conscience in general. In a passage quoted by Bennett he has Huck say

> It don't make no difference whether you do right or wrong, a person's conscience ain't got no sense, and just goes for him anyway. If I had a yaller dog that didn't know no more than a person's conscience does, I would pison him. It takes up more room than all the rest of a person's insides, and yet ain't no good, nohow.[32]

Those sentiments are very close to Twain's own. 'Mine was a trained Presbyterian conscience', he wrote 'and knew but one duty—to hunt and harry its slave upon all

---

[29] (Twain, 1884) Chapter xvi. The issue comes back again, especially in Chapter xxxii.

[30] See for instance (Driver, 2001) pp. 51–5; (Arpaly and Schroeder, 1999) p. 162; (Arpaly, 2003) pp. 75–9. Two authors offering interpretations closer to mine, though for rather different reasons, are (Freedman, 1997) and (Hirsthouse, 2001) pp. 150–3.

[31] The amended pages are reproduced in the Blair and Fischer edition of *Huckleberry Finn* that I reference.

[32] Chapter xxxiii.

pretexts and on all occasions; particularly when there was no sense or reason in it.'[33] And the same idea came up in many other places. For instance:

> All consciences I ever heard of were nagging, badgering, fault-finding, execrable savages! Yes; and always in a sweat about some poor little insignificant trifle or other—destruction catch the lot of them I say! I would trade mine for the small-pox and seven kinds of consumption, and be glad of the chance.[34]

This general suspicion of conscience, on both Twain's part and Huck's, makes better sense of Huck's attitude when he returns to Jim, having failed to turn him in:

> I got aboard the raft, feeling bad and low, because I knowed very well I had done wrong, and I see it warn't no use for me to try to learn to do right; a body that don't get started right when he's little, ain't got no show - when the pinch comes there ain't nothing to back him up and keep him to his work, and so he gets beat. Then I thought a minute, and says to myself, hold on—s'pose you'd a done right and give Jim up; would you feel better than what you do now? No, says I, I'd feel bad—I'd feel just the same way I do now. Well, then, says I, what's the use you learning to do right, when it's troublesome to do right and ain't no trouble to do wrong, and the wages is just the same? I was stuck. I couldn't answer that. So I reckoned I wouldn't bother no more about it, but after this always do whichever come handiest at the time.

Bennett interprets this as a rejection of morality altogether, and whilst there is something to this, it is only a rejection of morality in the rather narrow sense it which it is equated with the current dictates of his conscience (and that in turn in Huck's mind is traced to the lessons of people like Miss Watson, Jim's owner[35]). Since, at least in retrospect, Huck is aware that his conscience will disapprove whatever he does, he is aware that it is not a reliable guide to what he ought to do. Indeed, he is aware of this even as, setting out to tell on Jim, he is disturbed by Jim's cries of confidence and affection. In the published version he says:

---

[33] (Twain, 1906) quoted in (Hearn, 2001) p. 155. See also Twain's comments on conscience in *What is Man?* quoted in the same place. So I think that Thomas Hill is quite wrong when he voices his suspicion that 'Mark Twain had his tongue in his cheek when he attributed to 'conscience' Huck Finn's 'guilty' feelings about helping the slave, Jim to escape'. (Hill, 1998) p. 291. It is exactly because he thought that conscience was really at work here that Twain thought Huck's attitudes explicable.

[34] (Twain, 1876). See also (Kaufmann, 2006), for a list of some other occurrences of the same idea.

[35] Huck is, moreover, well aware that Miss Watson's morality is not the only one; see his discussion of the 'two Providences'—that of Miss Watson and that of the widow Douglas—in Chapter III.

> I went along slow then, and I warn't right down certain whether I was glad I started or whether I warn't.

In the revised reading copy, Twain makes clear that this is not just an issue of whether he was glad, but of what he thought he ought to do, adding:

> It kind of all *unsettled* me, and I couldn't seem to *tell* whether I was doing *right* or doing *wrong*.

And in yet later revisions, Twain shows that Huck thought his conscience was no use at all as a source of moral insight:

> I don't want no such thing around as a conscience… You ain't wanted, you ain't welcome, you ain't no use to me. I never see such a low-down troublesome cuss, I says. It don't make no difference what a person does, you ain't ever satisfied and you is as free as if you owned the whole layout. If I'd a give Jim up you' d a kep me awake a week mournin' about it; and now you're gittin' ready to try to keep me awake another week because I *didn't* give him up … I wouldn't be as ignorant as you for wages. You don't know right from wrong, you ain't got no judgment, you ain't got no sense about anything—you ain't no good but just to lazy around, find fault and keep a person in a sweat.[36]

Assuming that Twain is here not revising, but rather expanding, the novel, how should we understand Huck's behaviour? I think that there is little doubt that he shows weakness of will. Indeed, he plausibly shows it twice over. For, on first meeting the fleeing Jim Huck promises not to tell on him: 'I said I wouldn't and I'll stick to it. Honest *injun* I will. People would call me a low down ablitionist and despise me for keeping mum—but that don't make no difference. I ain't agoing to tell, and I ain't agoing back there anyways'. So in breaking his promise—a promise of which Jim reminds him as he paddles off to tell—there is reason to think that he has shown weakness of will. That, or course, is controversial, depending on whether we think that the pangs of conscience make it reasonable to revise his earlier resolution. Less controversial is the case for saying that he shows weakness of will when he fails to turn Jim in. Asked by the slave hunters whether Jim is white or black, he fails to go though with his resolution:

> I didn't answer up prompt. I tried to, but the words wouldn't come. I tried, for a second or two, to brace up and out with it, but I warn't man enough—hadn't the spunk of a rabbit. I see I was weakening; so I just give up trying, and up and says: 'He's white.'

[36] Cited in (Campbell 1992). It is unclear whether Twain ever read from this text.

Huck thinks he has shown weakness of will, and I think we should agree with him; admirable though his inability is, it is hard to see the revision as stemming from any new information or insight of the kind that should rationally have him reconsider. I think we should probably conclude that is a revision that, by the standards of resolution maintenance, he should not have made. The standards of decency and humanity say otherwise; but they are not what are in play when we are assessing weakness of will.

But if Huck is weak-willed, is he akratic? That depends on whether, at the time has acted, he believed that turning Jim in was the right thing to do. And I think that we have plenty of evidence to think that the answer to that is far from clear. He certainly thinks that his conscience is telling him that turning Jim in is the right thing to do; and there are time at which he seems to believe it. Equally though there are times when he doubts it. Admittedly, as Bennett says, the reasons he lists are all on the side of turning Jim in; but that is surely because his settled plan had for so long been to help Jim escape that his conscience had time to come up with arguments. Had his settled plan been to turn him in, and had his conscience indeed given him a hard time over that, there is little doubt that it would have marshalled reasons on the other side: that he had promised Jim, that Jim was trusting him, that Jim was a friend.[37]

So what should we say? We might say that Huck believes different things at different times; but the times are so close together that even that seems rather misleading. Or we might say that he has contradictory beliefs, or that he thinks himself in a moral dilemma, confronted with contradictory obligations. I would be more inclined to say that he doesn't have any clear and straightforward beliefs in any direction; he is in a quandary, pulled in different directions at the same time.

Does this matter? If Twain failed to write the book in which Huck is clearly akratic, couldn't we just imagine another account in which he is? I don't think that things are so easy. As Arpaly points out, it matters that it isn't just squeamishness that prevents Huck from turning Jim in.[38] It matters that he is acting, as she puts it, from something like a visceral sense of equality. Now it is conceivable that he could have such an attitude in the complete absence of any belief that what he is doing is right. Arpaly thinks that he does: 'he does not have the belief that what is does is right *anywhere* in his head'. But since even such an unreflective creature as Huck does, as we have seen, try to make sense of what he is doing, that seems unlikely. And he has a lot to make sense of: he does not simply to fail to turn him in, but he goes on to construct an elaborate and effective story about smallpox to keep the slave

[37] Bennett speculates that in Huck's morality 'promises to slaves probably don't count'. I don't see any reason for believing that. The chapter in question comes directly *after* the chapter in which Huck apologies to Jim for playing tricks on him, and effectively vows not to play any more.

[38] (Arpaly, 2003) p. 76. I don't think though that she is quite fair in attributing to Bennett the view that Huck is merely squeamish.

hunters from investigating the raft where Jim remains. His conscience certainly gives him a hard time over this, but as we have seen, he does not cede the argument to it. More likely that his earlier sentiment remains: 'I couldn't seem to *tell* whether I was doing *right* or doing *wrong*'.

Arpaly's description is much more plausible if applied to Huck before he came to act. Huck did start out with the clear belief that turning Jim in was right; resolving to act on that belief gave him a rare period of equanimity. 'I felt easy and happy, and light as a feather, right off. All my troubles was gone'. And Twain adds, in the reading version: 'O! it was a *blessed* thought! I never can tell how good it made me feel—cuz I *knowed* I was doing *right* now'. But Huck's confidence in that belief did not survive the attempt to act upon it. Here again we have an example of the phenomenon discussed before: a case of one's judgments being formed in the light of one's decisions, rather than existing prior to them and persisting through them.

I suggest that Huck is typical. It might seem a bit of a stretch to rest so much on one example, and I do not want to argue that cases of inverse *akrasia* are impossible. Indeed I am confident that plenty have been actual. But there are two forces that will make it rare. First, as in the case of standard temptation, there is simply the pressure of something like desire: Huck wants to protect Jim, and as he comes to realize that he will he will act on this desire, so he will come to shift his judgment on what is desirable. The second feature is harder to describe. In so far as what moves the agent is not desire, but the pressure of the reason to do what is right, however hazily that is apprehended, the agent will still be powerfully motivated to make sense of it. This need not consist of a shift to believing that the action is right. It is possible that it could be understood as succumbing to the pressure of evil. There are plenty of passages in which Huck provides this interpretation of his own actions, though as I have argued, I think he only sees himself as rejecting Miss Watson's morality, and he doesn't set much store by that. But in general, given how strongly motivated we are to arrive at an account of ourselves that puts us in a good light, it will be surprising if the thought that we are doing what is right does not get at least a toehold.

REFERENCES

Anthony, Louise 1993: 'Quine as Feminist: the Radical Import of Naturalized Epistemology', in Louise Anthony and Charlotte Witt (edd.) *A Mind of One's Own* (Bolder: Westview Press) pp. 185–225.

— 2000: 'Naturalized Epistemology, Morality and the Real World', *Canadian Journal of Philosophy* Supp. Vol. 26, pp. 103–37.

Arpaly, Nomy 2000: 'On Acting Rationally Against One's Best Judgment' *Ethics* 110, pp. 488–513.

— 2003: *Unprincipled Virtue* (New York: Oxford University Press).

Arpaly, Nomy, and Timothy Schroder 2000: 'Praise, Blame and the Whole Self', *Philosophical Studies* 93, 161–88

Baumeister, Roy 1996: *Evil* (New York: W. H. Freeman).

Bennett, Jonathan 1974: 'The Conscience of Huckleberry Finn' *Philosophy* 49.

Bratman, Michael 1987: *Intention, Plans, and Practical Reason.* Cambridge, MA: Harvard University Press.

— 1998: 'Toxin, Temptation and the Stability of Intention', in Jules Coleman and Christopher Morris (edd.), *Rational Commitment and Social Justice* (Cambridge: Cambridge University Press) pp. 59–83; reprinted in (Bratman, 1999) pp. 58–90.

— 1999: *Faces of Intention.* (Cambridge: Cambridge University Press).

— 2006: 'Temptation Revisited' in *Structures of Agency* (New York: Oxford University Press) pp. 257– 82.

Brehm, Jack 1956: 'Postdecisional changes in the desirability of alternatives' *Journal of Abnormal Psychology* 52 pp. 384–9.

Campbell, Gregg 1992: '"I Wouldn't Be as Ignorant as You for Wages": Huck Talks Back to His Conscience', *Studies in American Fiction,* Autumn 1992.

Carver, Charles and Michael Scheier, 1998: *On the Self-Regulation of Behavior* (Cambridge: Cambridge University Press).

Cooper, Joel 2007: *Cognitive Dissonance: Fifty Years of a Classic Theory* (London: Sage).

Driver, Julia 2001: *Uneasy Virtue* (Cambridge; Cambridge University Press).

Freedman, Carol 1997: 'The Morality of Huck Finn' *Philosophy and Literature* 21 pp. 102–13.

Hearn, Michael (ed.), 2001 *The Annotated Huckleberry Finn* (New York: W.W. Norton).

Hill, Thomas, 1998: 'Four Conceptions of Conscience' reprinted in *Human Welfare and Moral Worth* (Oxford: Oxford University Press 2002) pp. 277–309.

Hursthouse, Rosalind 2001: *On Virtue Ethics* (Oxford: Clarendon Press).

Holton, Richard 1999: 'Intention and Weakness of Will' *Journal of Philosophy*, 96, pp. 241-62.

— 2003: 'How is Strength of Will Possible?' in Sarah Stroud and Christine Tappolet (edd.) *Weakness of Will and Practical Irrationality* (Oxford: Clarendon Press) pp. 39-67.

— 2004: 'Rational Resolve', *Philosophical Review* 113, pp. 507-35.

— 2006: 'The Act of Choice', *The Philosophers' Imprint* 6, 3.

— 2009: *Willing, Wanting, Waiting,* (Oxford: Clarendon Press).

Jones, Karen, 2003: 'Emotion, Weakness of Will, and the Normative Conception of Agency', in Anthony Hatzimoysis (ed.) *Philosophy and the Emotions* (Cambridge: Cambridge University Press) pp. 181–200.

Karniol, Rachel and Dale Miller, 1983: 'Why not wait? A cognitive model of self-imposed delay termination'. *Journal of Personality and Social Psychology* 45, pp. 935–42.

Kaufman, Will 2006: 'Mark Twain's Deformed Conscience' *American Imago* 63, pp. 463-478

McIntyre, Alison 1990: 'Is Akratic Action Always Irrational?', in Owen Flanagan and Amelie Rorty (edd.) *Identity, Character, and Morality,* (Cambridge MA: MIT Press) pp. 379–400.

Mele, Alfred, forthcoming: 'Weakness of Will and Akrasia' ms.

Mischel, Walter 1996: 'From Good Intentions to Willpower' in Peter Gollwitzer and John Bargh (edd.) *The Psychology of Action,* (New York: The Guildford Press) pp. 197–218.

Nietzsche, Friedrich, 1886: *Beyond Good and Evil.* Trans. R. Hollingdale (Harmondsworth: Penguin, 1973).

Scanlon, Thomas. 1998: *What we Owe to Each Other* (Cambridge MA: Harvard University Press).

Schelling, Thomas 1960: *The Strategy of Conflict,* (Cambridge MA: Harvard University Press).

Twain, Mark 1876: 'The Facts Concerning the Recent Carnival of Crime in Connecticut' *The Atlantic Monthly,* June 1876, pp. 641–50.

— 1884: *Huckleberry Finn,* edited by Walter Blair and Victor Fisher (Berkeley and Los Angeles: University of California Press, 1988)

— 1906: 'Chapters from my Autobiography', *The North American Review* September 7 1906.

Wang, Jing, Nathan Novemsky, Ravi Dhar, and Roy Baumeister, Forthcoming: 'Effects of Depletion in Sequential Choices' ms.