IV Strength & Weakness of Will

WEAKNESS OF WILL

If Akrasia is intentionally acting against one's best judgment, then there should be a distinct phenomenon of over-readily revising an intention; and a more restrictive notion of over-readily revising a resolution, where a resolution is thought of as an intention that is meant to be proof against later temptation.

GEORGE AINSLIE'S EIGHT MARKS OF THE WILL

1. A new force in addition to desires etc.
2. Throws its strength on the weaker side. (or perhaps it *can*)
3. Unites actions under a common rule. (?)
4. Strengthened by repetition
5. Very vulnerable to nonrepetition
6. Requires no diversion of attention
7 Doesn't normally depend on each single choice, except where stakes are very high.
8. Tendency of failure in one sphere to spill into others is variable.

WHY DO WE EXHIBIT WEAKNESS OF WILL?

Exponential and Hyperbolic discounting. $100 now or $120 in a year's time; $100 in ten year's time or $120 in eleven year's time. From pigeons to people (though non-humans are not good at delaying reward). Hyperbolic discounting is, in some sense, irrational.

Hyperbolic discounting is probably also involved in what we might call *one more won't hurt* reasoning:

> *One more won't hurt*
> You are a smoker, currently smoking forty cigarettes a day. You believe that this is very bad for you, and that you should give up. But you reason as follows: one more cigarette won't hurt me, and it will give me considerable pleasure now. So I might as well have this one more. And then you use this argument forty times a day.

Without hyperbolic discounting it looks as though the subject would simply be making a mistaken calculation (is that so in all cases?); but with it they need not be. Is hyperbolic discounting an *explanation* or merely a *description*?

AINSLIE'S EXPLANATION OF HOW STRENGTH OF WILL IS POSSIBLE

Competition between egoistic time slices, akin to an interated prisoners' dilemma. ('Picoeconomics') The reward game. See how this gives us the eight marks of the will.

PROBLEMS WITH AINSLIE'S EXPLANATION

(i) General problem of time slices: how short should they be? Can we make sense of actions done by time slices?

(ii) Each time slice is dead by the time of fulfillment, so in what sense is there a long run benefit for which they can trade?
(iii) What has become of the irrationality?

REVISING AINSLIE'S EXPLANATION

An alternative, solving the second problem: give up on the egoism. Assume that the time slices have desires (i) for others' well being, which (ii) they will never know to have been fulfilled. But:

(a) we've now lost the sense of competition; this is quite unlike a prisoner's dilemma.
(b) even these altruistic desires aren't enough, since they don't explain a further phenomenon, that having kept to a resolution provides a reason for continuing to do so (compare the sunk cost phenomenon).
(c) it still doesn't help with the third problem. Why should we think that it is irrational for different actors to have different desires?

A MORE RADICAL REVISION

Better just to put back the enduring self. What is left from the prisoners' dilemma model is just the idea that in deciding what to do, we are acutely sensitive to what else will be done, but by us at a later time, rather than by others. This still keeps all of the advantages. In particular, we keep (indeed can make even better sense of) this reasoning:

(i) If I resist temptation now, then I'll resist it in the future.
(ii) If I don't resist temptation now, then I won't resist it in the future.

OTHER FEATURES OF THE PROPOSAL

The status of rationalizations and exceptions.
Bright lines. Alcoholics Anonymous on the need never to drink.

WORRIES

(i) Does this help explain all cases of strength of will in the face of desire? What about a case in which there is no likelihood of repetition? The highly unusual one night stand.
(ii) Aren't the conditionals implausible? (Most people have several goes at stopping smoking.)
(iii) Does this explain how there can be *effort* involved in resisting temptation?

EGO-DEPLETION

Baumeister *et al.*: it seems that the will does resemble a muscle. Prior resolutions seem to prevent ego-depletion. (Many studies *seem* to show this, but there are now some problems with replication.)

SEPARATE NEURAL SYSTEMS: MCCLURE *ET AL.*

Decisions involving immediate reward activate parts of the limbic system associated with the midbrain dopamine system. In contrast, intertemporal choices activate regions of the lateral

prefrontal cortex and posterior parietal cortex. Deciding to delay brings greater activity in these latter regions.

NORMATIVE CONCERNS

Could the two conditionals ((i) and (ii) above) give me *information* about what I am like? Problem: is it accurate information?

An extreme case: Newcomb's problem. A bizarre billionaire offers you a choice:

(1) Box A
(2) Box A *and* Box B

Box B contains $1000, placed there the night before. Box A either contains $1m, placed there the night before, or else nothing. The billionaire is a brilliant predictor of people's choices (with a 99.9% success rate). When he decided last night what to place in Box A, he contemplated whether you would choose one box, or both. If he thought that you'd greedily choose both, he placed nothing in Box A. If he thought that you would choose only one, he placed $1m in Box A. But that, of course, was yesterday, and nothing you can do now will affect what is in the boxes. Should you choose one box or two?

Less extreme: self-signaling behavior. Hardworking Calvinists. People keeping their hands in cold water for longer if that indicates a strong heart (Quattrone and Tversky, 1984). The difficulty is that if they see themselves as self-signaling, they interfere with the very signals that they think are giving them information. But in the temptation resisting case does that matter? There isn't any *further* state that the behavior is supposed to be evidence of. If someone behaves apparently altruistically because they want to think of themselves as an altruist, then maybe they are not truly an altruist. But if someone refrains from smoking because they want to think of themselves as someone who can refrain from smoking, then they really are refraining from smoking. Compare simple self-presentation behavior.

TWO OTHER PUZZLE CASES

The reciprocal suitcase deal (compare Broome's wolf case).
The self-torturer (note that this case, like that of Ann, seems to involve the idea not simply that desire changes as time changes, but that it changes in response to other behavior).

BRATMAN'S TROUBLEMAKING PRINCIPLES

The Linking Principle: I shouldn't form an intention that I now believe I should, at the time of action, rationally revise.

Or, more precisely:

If, on the basis of deliberation, an agent rationally settles at t_1 on an intention to A at t_2 if (given that) C, and if she expects that under C at t_2 she will have rational control of whether or

not she A's, then she will *not* suppose at t_1 that if C at t_2 she should, rationally, abandon her intention in favor of an intention to perform an alternative to A. (Bratman, 1998, 64)

The standard view: My ranking now should depend on the ranking that I will make at the time.

TWO PUTATIVE SOLUTIONS

(i) *Sophistication*: accept linking principle, and standard view. Problem: you'll give into temptation

(ii) *Strong resolution*: accept linking principle, but reject standard view, in favor of the idea that one's ranking now should not depend upon the ranking that one would give at the time of action, but instead at the time of forming the resolution. Problem: seems to make too little of our agency at the time of action. (More like a machine that is locked into the plan.)

BRATMAN'S SOLUTION

Add another condition onto what is needed for rationality: you must meet the *no regret condition*:

(i) were you to stick with the resolution, then at plan's end you would be glad about it; *and*

(ii) were you to fail to stick with it, then at plan's end you would regret it.

Does this help with the toxin case, and the other cases?

THE CASE OF YURI

Yuri has managed to fall in love with both Tonia and Lara. When he is with Tonia he is convinced that she is the one, and vows his undying commitment; unfortunately things are just the same when he is with Lara. Worse still, his life is so structured that he keeps spending time with each of them. As one commitment is superseded by another, and that by another, trust is lost all round. Clearly it would be rational for Yuri to persist in his commitment to one of the women, and to restructure his life accordingly; all of them recognize that. However, the no-regret condition isn't met. We can imagine him as a naturally contented type, who will not feel regret whomever he ends up with; in which case the second clause of the condition would not be met. Or we can imagine him as a naturally discontented type, who will feel regretful either way; in which case the first clause will not be met. Or we can imagine him as ambivalent, fluctuating between regret and happiness however he ends up; in which case neither clause will be stably met.

TWO VERSIONS OF THE LINKING PRINCIPLE AND A PRINCIPLE LINKING THEM

Weak Link: I shouldn't form an intention that I now believe I should, at the time of action, rationally reconsider and revise

Strong Link: I shouldn't form an intention that I now believe that if I were, at the time of action, to reconsider, I should rationally revise

Rational Reconsideration Principle: If I now believe that if I were to reconsider at the time of action I would reasonably revise, then I should reconsider at that time.

A COUNTEREXAMPLE TO STRONG LINK

You are defending your ship. Your instruments tell you that you are being attacked from somewhere in a 30° arc to the North East. If you waited and calculated you could find out the exact position of the attacker. But you are anticipating further attacks that will need your attention. Rather than waiting, finding the exact position of the attacker, and responding with a single missile, you form the intention of launching, when the optimum time comes, a barrage of missiles to cover the whole arc. In effect you trade missiles for time to attend elsewhere.

THE RATIONALITY OF STRONG RESOLUTION

There is a difference between (i) reconsidering a resolution, deciding that it would be rational to revise and then not revising; and (ii) not reconsidering even if, were you to reconsider, you would think it rational to revise. The former is irrational; the latter is not. Does this make rationality too fragile? Not if reconsideration is an involved business. Perhaps you only come to a judgment about what is best subsequent to the formation of an intention

EXPLAINING WHAT IS RIGHT ABOUT THE NO-REGRET CONSITION

Regret as *indicating* substantial mistake concerting one's choice (not *constituting* formal mistake).

BACK TO KAVKA AND THE OTHER PARADOXES

Maybe we are in fact not so constituted that we could form the intention; but that doesn't show that it would be irrational. Consider how we would bring up in children in a world in which there were many toxin cases.

Now think about how we morally bring up children. We teach them not to reopen the question of what to do once they have promised. And more generally we teach them not to reopen the question of what to do once they have knowingly induced someone to rely on them. Perhaps this is a central moral skill.

What about Newcomb cases? There things are somewhat different, in that there is no prior commitment. But suppose we got people to commit to a Newcomb policy; and suppose they committed to one-boxing. Would they then be irrational if they failed to reconsider that policy when the time came to act?