LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

**Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin**

**Copyright statement**

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

**Liability statement**

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

**Access Agreement**

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

# Weakness of Will and Practical Reason

A dissertation submitted to the University of Dublin, Trinity College,
For the degree of Doctor of Philosophy

by

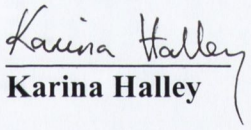
Karina M. Halley M.A. (research)
Department of Philosophy
Trinity College, Dublin

January 2004

TRINITY COLLEGE
0 2 JUN 2004
LIBRARY DUBLIN
THESIS
7952
7491

## Declaration

I declare that this thesis is entirely my own work except where acknowledged. This thesis has not previously been submitted for a degree at this or at any other university. I agree that this thesis may be lent or copied at the discretion of the Librarian, Trinity College, Dublin.



**Karina Halley**

**Summary**

This thesis argues that a correct understanding of weakness of will (that is, freely failing to act as one thinks one has most reason to act) is crucial to a correct understanding of practical reason. Central to a theory of practical reason is the question of how practical reasons are related to motivation and action. On a common view (the Internalism Requirement), if an agent thinks she has reason to act, then *ceteris paribus,* she will act. However, I argue that this view is difficult to reconcile with the possibility of weakness of will. Indeed, a proper examination of weakness of will can serve to determine to what extent the Internalism Requirement is acceptable and to what extent it needs to be revised or even abandoned.

In Part I I examine strategies for reconciling the internalism requirement with weakness of will. These can be broken down into two broad positions on practical reason, strong internalism and weak internalism. I argue that neither position is capable of making the phenomenon of weakness of will consistent with the internalism requirement. Strong internalism argues for a necessary connection between having a reason, all-things-considered, and being motivated to act accordingly. The problem with this view is that if we take into account weak-willed action as described then we are either forced to reject this strong version of internalism altogether or deny that weak-willed actions *as described* actually occur. Weak internalism argues that the connection between having a reason and being motivated to act on it is a defeasible one; that is, insofar as an agent is rational, she will not fail to act on an all-things-considered reason. The problem with weak internalism is that it fails to show that it is necessarily impossible even for an agent qua rational to have a normative reason to act without the motivation to act. I argue further that both strong and weak internalism

seem to beg the question regarding the nature of the connection between having a reason and being motivated to act on it.

In Part II I further examine the weak internalist strategy insofar as it is adopted by neo-Kantians who appeal to the existence of irrational action generally to strengthen the case for a theory of pure practical reason (that is, the view of practical reason whereby reasons themselves are motivational and require no extra connection with independently intelligible motives for action). On this neo-Kantian view weakness of will and other irrational actions are simply examples of conditions that can cause us to act irrationally, but such actions can no longer be used to deny that there must be principles of conduct other than hypothetical imperatives. Furthermore it is argued that weakness of will cannot even be explained on the opposing Humean view of practical reason as instrumental. I examine the considerations that lead neo-Kantian theorists to hold such a view before going on to argue that those favouring the pure practical reason view end up begging the question regarding weakness of will because of their view that to endorse a reason for action is to evaluatively concludes that that reason is best, all-things-considered.

In Part III I attempt to tell a somewhat different story about the relation between endorsing a reason and acting on it. The neo-Kantian view of what it is to endorse a reason is one I suggest we should reject and replace with a different understanding. What is disputed is the neo-Kantian view that endorsement amounts to evaluative judgement. I suggest instead that *attitudes* play a crucial role in endorsement. By arguing that we can explain cases of weakness of will more successfully in terms of attitudes and their stability we show that the neo-Kantian criticisms do not apply to properly developed Humean or naturalistic views. Thus the way lies open to further develop this attitudinal account of commitment to action into a full-blown theory of practical reason.

**Weakness of Will and Practical Reason**

Karina M. Halley

Abstract

How are we to characterise the reliability of the connection between having a reason to act and being motivated to act? This question is central to a theory of practical reason. This thesis argues that an examination of weakness of will is crucial to answering this question. In particular, examination of this issue is crucial to trying to resolve the dispute between a Kantian view, which argues that practical reasons are by themselves motivating, and a broadly Humean view, which denies this. It is argued that arguments against Humean-inspired accounts of practical reason on the basis that they cannot account for weak-willed or irrational action are problematic. Consequently they do not give us sufficient reason to reject a broadly Humean account of practical reason. A sufficiently sophisticated Humean view can indeed account for practical reasoning. Neo-Kantian theorists argue that this is not possible. However by outlining an alternative view of how practical reasoning actually works it is suggested that we can both satisfactorily account for weakness of will and avoid certain difficulties that a neo-Kantian account of practical reasoning leads us into.

**Weakness of Will and Practical Reason**

Karina M. Halley

Abstract

How are we to characterise the reliability of the connection between having a reason to act and being motivated to act? This question is central to a theory of practical reason. This thesis argues that an examination of weakness of will is crucial to answering this question. In particular, examination of this issue is crucial to trying to resolve the dispute between a Kantian view, which argues that practical reasons are by themselves motivating, and a broadly Humean view, which denies this. It is argued that arguments against Humean-inspired accounts of practical reason on the basis that they cannot account for weak-willed or irrational action are problematic. Consequently they do not give us sufficient reason to reject a broadly Humean account of practical reason. A sufficiently sophisticated Humean view can indeed account for practical reasoning. Neo-Kantian theorists argue that this is not possible. However by outlining an alternative view of how practical reasoning actually works it is suggested that we can both satisfactorily account for weakness of will and avoid certain difficulties that a neo-Kantian account of practical reasoning leads us into.

## Acknowledgements

I have benefited immensely from the help and support of many friends and colleagues during my time at Trinity. I would especially like to thank Dr Vasilis Politis for his excellent supervision, generosity of time, and confidence in me over the years. I would also like to thank Trinity College, Dublin and Atlantic Philanthropies for their valuable financial support. The Department of Philosophy has provided me with a congenial and supportive environment over the years and for this I am most appreciative. I would like to express my gratitude to my family for their encouragement and support over more years than I can remember. Finally, I would like to thank my husband, Jim, for his never-failing confidence and pride in my work, and for his continued patience and understanding.

**Weakness of Will and Practical Reason**

## Contents


**Introduction** 1

**Part I** 24

**1 Weakness of Will and Motivational Internalism (1): The Problems** 26

1.1 Weakness of will and the explanation of intentional action: a dilemma 34
1.2 Action and motivation: an outline of the internalist perspective 47
1.3 The rise of internalism: a brief sketch 48
1.4 Internalism and the explanation of weak-willed action 55
1.5 Strong internalism 57
1.6 Weak internalism 64

**2 Weakness of Will and Motivational Internalism (2): The Solutions** 70

2.1 The 'usurping reasons' view 72
2.2 The 'no genuine judgement' view 83
2.3 The weak internalism solution to the problem of weakness of will 89
2.4 What weak internalism really commits us to 95

**Part II** 111

**3 Pure Practical Reason and the True Irrationality View** 117

3.1 Motivational scepticism 117
3.2 The instrumental principle: norm or description? 121
3.3 The necessity of rational requirements 139
3.4 The IR view and irrational action 151
3.5 'Mere' descriptions 159
3.6 The IR agent: a jumble of desires? 161

**4 Weakness of Will and Endorsing a Reason for Action: PPR Criticisms Examined** 167

4.1 The free rider 169
4.2 The deliberative agent 173
4.3 Deliberative experience 189

**Part III** 207

**5 Practical Reasoning and Intending to Act** 209

5.1 More internalist assumptions 209
5.2 Must one intentionally act as one intends? 214
5.3 The normative 'gap' and norm-governed transitions in practical reasoning 224
5.4 Phronesis and commitment 228
5.5 Sustaining commitment: stability 237
5.6 Weakness of will and the authority of practical reason 244

**Conclusion** 249

**Bibliography** 255

## Introduction

### What is the problem of weakness of will and why is it relevant?

When we think about any kind of action there is one prominent intuition we find impossible to do without, and that is that actions always seem to follow upon some sort of *judgement* or consideration (understood in the weakest and widest possible sense). We report this introspectively and it is something we assume everyone does as they navigate their way through everyday actions. Whether people act rashly, unwisely or in a way that meets general approval and admiration we firmly believe that at least in some minimal sense they "know what they are doing" because it just does not seem conceivable that *action* could occur without the agent who is doing the acting *directing* what she does. If we tease out this intuition further what is at the bottom of all this is the view that we are somehow in control of our own actions. I 'decide' what I want to achieve, I 'set my priorities' and I 'figure out' how best to go about things. The overall picture is of an agent who has executive control over her action. It is obvious, therefore, that any account of practical reasoning (that is, reasoning about what to do) will have not merely to explain why this intuition is so powerful, but must treat it as a phenomenon to be saved.

A bit more formally, we can express this intuition as what can be called an *Internalism Requirement* on action. We think it no accident that people tend to do what they take themselves to have good reason to do and we believe that:

If an agent thinks she has reason to $Q$, then *ceteris paribus*, she will $Q$.

This really does no more than convey our ordinary intuition. We do not believe that agents will always do whatever they see themselves as having reason to do because we just do not expect either the strength of will to be available or we are mindful of being in situations where a decision has to be made between competing reasons for action.

We can fit a lot of mitigating circumstances into the *ceteris paribus* clause of the internalism requirement. No one would suggest that we can conclude that our intuitions about action are wrong merely because agents sometimes fail to act according to a self-addressed reason for action. If, for example, Mary thinks she has a reason to turn on the radio (say, because she expects an important announcement at six o'clock) but fails to do so, can this act as some sort of counter-example to our intuition about action always following upon the considerations of the agent? In one sense obviously not. If Mary is tied to a chair at the time or is completely paralysed from the neck down then there is no conceivable way she can turn on the radio, no matter what reasons she may have about doing so. But in another sense there are many questions left unanswered here about what genuinely excuses an agent for failing to act on her reasons. Why are we more willing to excuse Mary for not turning on the radio if she is tied up than if say, she admits she has compelling reason to do so but says she just could not be bothered? In the former case she is willing but not able whereas in the latter she is able but not, in some sense, willing. Why do we look upon the latter case with disapproval and on the former with pity or understanding? We can be in no doubt that there are normative standards governing what counts as an excuse for inaction and we will be meeting this justificatory question again later.

The problem of interest to us here however concerns what appears to be an obvious and difficult counter-example to the internalism requirement, and that is this: how are we to understand situations where there are no obvious impediments to action (i.e. where all things *are* equal) and yet the agent fails to act? Consider Mary again. Imagine she thinks that she has good reason to turn on the radio (she wants to hear the announcement). She is not incapacitated in any way but nevertheless she just fails to turn on the radio. Is this a case of what we call *weakness of will*? Well not necessarily, because Mary may have had other reasons to act. Maybe she thought she

had a reason to turn on the radio and yet another reason to turn on the television. In this situation Mary has *prima facie* reasons for action – she has a reason to do each action and until she decides which reason to act on *all things considered* she can be said to have reasons which are *prima facie* or prior to further consideration. Mary's failure to turn on the radio only seems weak-willed and puzzling before we found out that she may have had a stronger reason in favour of doing something else. Now imagine that upon examination of Mary's situation we discover that she in fact has no other, stronger reason overriding her reason to turn on the radio, neither is she incapacitated or anything like it, and yet she still fails to turn on the radio to hear the announcement. How should we interpret Mary's lack of action? The common answer here is that she is 'weak-willed' – she does not have the motivation to act as she judges best and so Mary is supposed to be to this extent *irrational* because she has good reason to turn on the radio and fails to do so despite that fact that having a reason to act is supposed to be in some sense motivational.

What we have here however is not so much an answer to our problem regarding how to characterise agents such as Mary, but a more elaborate description of the problem itself. To claim that an agent who fails to act as she thinks she should is weak-willed and that that entails being irrational because no rational agent would fail to be sufficiently motivated to act on a reason for action all things considered, raises more questions than it answers. To start with, by giving this explanation for Mary's inaction it is assumed that what it is to have a reason is to find oneself motivated to act accordingly.[1] Yet this is problematic precisely because it seems to make situations such as Mary's impossible. For she clearly does not find herself motivated to act and so we might be forced to conclude that she never really had any all-things-considered

[1] This does not mean that the agent *will* therefore act accordingly, but only that she will have the motivation to act should she eventually decide that this is what she ought to do *all-things-considered.* Remember the reason to act may only be a prima facie reason at this stage.

reason to turn on the radio in the first place. But this directly contradicts another aspect of our intuition about action, which is that agents can fail to act according to their all-things-considered reasons for action. If what it is to have a reason at all is to find oneself motivated to act accordingly then anyone who, like Mary, fails to act must not really have had a reason to act in the first place. If they never had any reason to act they can only be criticised for such things as saying they had a reason to act when they did not, culpable ignorance of the facts, or self-deception etc. Such an explanation is clearly contrary to our intuitions since we also believe that Mary *can* be criticised for not acting in this case and the form this criticism usually takes is criticism about her capacity to reason about what to do. Any agent who fails to take the means to her desired end we usually find at best puzzling and we accuse them of self-thwarting inconsistency.

Any explanation of failures such as Mary's will have to leave us room to criticise agents like this for being in some sense culpably deficient in a situation where they ought to recognise that having sufficient reason to act is in some way binding on them. What will become clear as we examine what is traditionally known as the problem of weakness of will is that this is no minor puzzle about practical reasoning, but instead cuts to the heart of the question of the very nature of practical reasoning itself. One problem we have witnessed is that it is very difficult to outline exactly what the problem is without committing oneself to a presupposed view of practical reasoning. What we are trying to capture is our intuition that action is somehow reliably linked to the holding of a judgement or reason in favour of that action (it is too early to declare whether by this we mean that action is *necessitated by* or *caused by* judgement). But we need to make this consistent with our other intuition that agents can and do fail to act according to their self-declared chosen judgement or reason for action. However, we do not want to make this intuition consistent with the previous one by redescribing

it in such a way that we end up arguing that agents who fail to act must be considered not to have had properly held judgements or reasons for action at all.

So we want to save what we might call our weakness of will intuition, but we have to be careful how we define the sort of action we are talking about. To say that there is something wrong with an agent who fails to be motivated by her all-things-considered reason for action presupposes that the *recognition* that one has an all-things-considered reason is sufficient to motivate an agent to act. But is this the intuition we want to save? If we held the view that such reasons, by themselves or the recognition that one has such a reason, cannot motivate but require a further 'push' from other contingent states of the agent such as her interests or desires, then an agent who fails to act as she has most reason to cannot be criticised solely on that basis. This agent would on such a view not be weak-willed unless she failed to act despite having the reason to act *and* having the relevant independently intelligible and contingent motivation. But if we instead hold the view that reasons are themselves motivational then it becomes more difficult to see how an agent who fails to act on an all-things-considered reason could really have had that reason in the first place. The former view broadly represents an *externalist* position on practical reason and the latter an internalist requirement on action.

Whichever view we take here affects not the basic intuition that sometimes agents fail to act on their all-things-considered reasons for acting but the *explanation* given for how this intuition can be made consistent with our view of practical reasoning. The apparent problem inherent in both these positions is the same – presented in these forms neither view can save the phenomenon of weak-willed action because on each view the agent in question is missing something crucial to practical reasoning about what to do. The externalist's agent in such cases does not have what is regarded as the independently intelligible motivator necessary to move her to act, so no

matter, for example, how much sense the agent thinks it makes to turn on the radio to hear the announcement, unless she has the inclination to do so she will not act and cannot reasonably be criticised for not doing so. The internalist's agent is similarly unmotivated to act but if she really had an all-things-considered reason to turn the radio on in the first place then the motivation and action should occur as a matter of course. Since it does not then we seem forced to conclude that she never had the all-things-considered reason in the first place, so she must be deceived in some way about what she has reason to do.

In these rather crude forms neither view can admit of the existence of a phenomenon we all experience from time to time – simply failing to act as we know or have decided that we ought. But any theory of practical reason has to preserve the idea that we can act irrationally, weakly, akratically or whatever we choose to call these phenomena. It is sometimes taken for granted that the problem we are dealing with here amounts to no more than what we are to call such a phenomenon in the sense that this involves distinguishing it from such things as *self-deception* or *wantonness,* but this approach belies a fundamental lack of appreciation for the true significance of the problem. We cannot approach the question of the nature of practical reason itself without first carefully considering the different ways in which we do act or how we *fail to act,* since to do so inevitably imports some considerable presuppositions regarding what it is to be a rational agent in the first place. It could be said that in order to show that weak-willed action is an example of 'irrational' action we might say that we need to know what it is to act rationally and so any view on the nature of irrational action must emerge out of a theory of practical reason based on the practical reasoning of the rational agent. Indeed it is difficult to see how we could do things differently. But the problem is that before any theory of practical reason is proposed we have already

separated out 'rational' actions from their 'irrational' counterparts and this is all too often where the story stops.

For an action to be rational usually means (broadly speaking) that it has been considered and judged appropriate in light of the agent's beliefs and goals. Given this we would assume that being *irrational* means that an act has not been so considered or has been found wanting with regard to its justification in some respect. But this is not the usual charge made in the case of irrational actions. Failure to act or *failure to be motivated* as one should is also submitted to the charge of irrationality. But then what exactly does it mean to say that an action is irrational or that an agent who fails to act is in some sense irrational? We usually take it to mean that the agent failed to direct her actions or her reasoning about what to do in a manner in accordance with practical reason.

What we need from a theory of practical reason is to be able to distinguish between action performed by an agent who has considered her reasons and acts accordingly, and action performed without such a proper reflective procedure. But how we determine what 'proper reflection' is in this context is crucial. If we want an accurate theory of practical reasoning then we have to start by considering all of the significant manners in which we classify actions in general; we cannot proceed by assuming that only certain actions are to be considered examples of properly thought-out actions because only these are rational. The issue of what sorts of actions and what sorts of agents are to be considered rational cannot be foreclosed before we start out on a theory of practical reason. If we are to start from the beginning, as it were, with all sorts of actions to be adequately explained by our theory of practical reason, then it is obvious from the outset that any view that makes weak-willed action impossible by giving it a redescription of whatever kind has to be considered unsatisfactory. This means, I suggest, that the internalism requirement cannot be interpreted as determining

that the agent who fails to act on her all-things-considered reason never really had that reason in the first place; and similarly, the externalist is not free to remove the motivational or directive force of reasons from reasons *qua reasons*.

Rather ironically it is only in the twentieth century that the problem of weakness of will became divorced from the larger question of the nature of practical reason. Historically the two have always gone hand in hand. Plato, for example, considered *akrasia,* as he called it, to be the result of an error or slippage in the proper functioning of the smooth control of the soul by its highest faculty, reason.[2] From the 1960's onwards the counter-Wittgensteinian rush to rediscover a causal account of action in terms of the agent's beliefs and desires and at the same time to give reasons pride of place in the etiology of action, resulted in a gradual shift of focus from metaphysical issues such as act-individuation[3] to a debate concerning why agents act for some reasons and not for others.[4] Puzzles thrown up by the requirement that agents always want to do and will do what they judge best to do generally centred on the attempt to make the existence of weak-willed action consistent with such a requirement despite the prima facie inconsistency. The central issue that emerges from the literature here is whether Davidson succeeds in ironing out the logical puzzles involved in the weak-

---

[2] Of course Plato's views on akrasia evolved between early dialogues such as the *Protagoras* and later works such as the *Republic.* It is plausible to see the former as an expression of a Socratic view and the latter as Plato's more considered later view. For Socrates, deliberate action *only* makes sense when understood as action in accordance with one's better judgement. To act against better judgement has to be rejected as ludicrous and impossible because the very ideas is a threat to what it means to act deliberately. It is clear that Socrates has in mind a very tight connection between knowledge and action: there is no room for further decision making once one knows what the right thing to do is. However what we find in later dialogues is a less sparse and more plausible understanding of motivational conflict which brings with it a new perspective on weak-willed action. Plato challenges the Socratic position when he accepts that a coherent account of deliberate action *can* incorporate the idea that agents *sometimes* knowingly act against their better judgement. However Plato accepts that a coherent account of deliberate action is to be understood as action performed according to the agent's best reasons or better judgement. For Plato, while reason can resist desire, it has to put up a fight. Socratic reason simply has to make a judgement and all the desires fall into place. Platonic reason is more like a preventative force, capable of resisting and preventing desire from taking the upper hand, but there is always a risk that the elements of the soul will "trespass on each other's functions." (*Protagoras, Meno*, 77a-78a; *Hippias Minor*, 373 *seq.*; *Gorgias*, 468b-500; *Republic,* bk IV; *Timaeus* 86d *Laws*, 860d-e, 864b-c.)

[3] Davis (1970), Ginet (1990), McCann (1972), Thalberg (1977), Thomson (1977).

[4] Anscombe (1963), Goldman (1970), Davidson (1963).

willed agent's failure to act according to her all-things-considered reason for action, given that all agents want to do what they judge best and reliably act accordingly. While Davidson's view marked an important turning point in developing a modern *naturalist* theory of practical reason it also served to mask certain fundamental issues which, when uncovered, expose the important normative questions for any theory of practical reason and intentional action.

Further on in our study, when we see the limitations of treating weakness of will as simply a puzzle for the logic of intentional action two central though competing claims emerge requiring examination:

1 We can appeal to the existence of irrational action to *deny* that pure practical reason [5] can by itself motivate the agent to do what she judges valuable or good.

2 We can appeal to the existence of irrational action to *strengthen* the case for pure practical reason.

We shall call views that tend towards (1) the Instrumental Reason View (IR for short), and (2) the Pure Practical Reason View (PPR). Both views are internalist in so far as advocates of each accept some version of the Internalism Requirement, but according to a traditional Humean or IR view, in order for some consideration to be a reason one must be favourably disposed to that consideration to begin with; while according to PPR acknowledging or recognising that one has a binding reason is by itself capable of motivating one to act. Those who accept a view that fits broadly with IR tend to see weakness of will as involving a conflict between the contents of a particular present desire an agent may have and a more general desire that the agent is supposed to be more persistently committed to on a long term basis. Those who, on the other hand,

[5] That is, the view of practical reason whereby reasons by themselves are motivational and require no extra connection, deliberative or otherwise, with independently intelligible motives for action.

accept a view that fits broadly with PPR tend to see weakness of will as involving a conflict between the contents of a particular desire the agent has and pure practical reason.

Most recent commentators attempt to provide a solution to the problem of weakness of will based on one of these general strategies, but again the bigger issues are in danger of becoming obscured. Christine Korsgaard, for example, puts forward an influential PPR view.[6] This view is that scepticism about whether pure practical reason can motivate us to act according to what it determines as valuable can only be based on scepticism about the capacity of practical reason to determine what is valuable in the first place. Once we see that scepticism about this special capacity of practical reason is unjustified then we will see that scepticism about the related motivational capacity of practical reason is not justified either. Korsgaard argues that we can hold both that irrational action can occur *and* that motivational scepticism has no force. If agents can be irrational despite the fact that they accept the truth of a moral (or a generally evaluative) judgement then there are no independent grounds for arguing that holding a moral (or evaluative) judgement is not a sufficient ground for action. Confusion on this point, she believes, has led people to seriously consider externalism. She thinks that by showing that the internalism requirement does not imply that agents *will* act on their judgement, and furthermore by establishing that reasons are not solely private states of agents, we will see that practical reason can indeed determine (through reflection) what is valuable. We *cannot* appeal to motivational scepticism (and weakness of will is often thought to be the most recognisable form of this) to deny that practical reason can determine what is worth valuing but we *can* appeal to weakness of will and irrational action generally to strengthen the case for an account of pure practical reason.

---

[6] Korsgaard (1987), also Hampton (1998), Nagel (1970).

On this view the phenomenon of weakness of will is supposed, somewhat ironically, to threaten naturalistic views on practical reason and make inevitable what Thomas Nagel would consider to be an 'agent-neutral' conception of practical reason, that is, the PPR view. Weakness of will cannot, according to Korsgaard, be understood as a conflict of desires in the manner in which those who advocate IR see it. To think of weakness of will as a conflict of desires will lead us to the absurd position where an agent's reason for action disappears as soon as she acts against it for a lesser reason; for "if this general desire for the good does not remain predominant, not only the motive, but the reason, for doing what will conduce to one's greater good, disappears."[7] Beating Hume at his own game, Korsgaard argues that any such view cannot account for weakness of will and so we should think of this phenomenon instead in terms of a conflict between pure practical reason and a particular desire. Of course along with this we must accept the inevitability of the PPR view.

I wish to argue however that these considerations favoured by Korsgaard and other non-naturalists fails to lead us necessarily or inevitably towards a pure practical reason view. It *is* possible, I want to argue, to explain weakness of will on the IR view which sees it as in some broad sense involving a conflict of desires. If this argument succeeds then weakness of will or irrational action generally cannot be used to strengthen the appeal to pure practical reason. However we can go further. Korsgaard and others who favour the PPR view, claim that *only* the PPR view can account for irrational action as well as intentional action generally. But if we can show that weakness of will can be explained on the view Korsgaard rejects, we deprive the PPR view of one of the primary advantages its advocates promote it as having over and against the IR view.

---

[7] Korsgaard (1986) p.15.

The outcome of these arguments will decide how we are to see practical reason. The role of weak-willed action in all this is central. If I am correct in arguing that what Korsgaard is calling 'motivational scepticism' actually *does* succeed in casting doubt on the capacity of pure practical reason to determine what is valuable, then allowing for the existence of weak-willed action in effect sets a test-case or precondition to be fulfilled on any possible theory of practical reason.

Davidson's solution to the problem of weakness of will is an attempt to investigate whether it is possible to come up with a solution consistent with the IR view. Weak-willed actions are the result of a conflict between a prima facie reason for action and an all-things-considered reason. The agent judges that *all things considered* it would be better *not* to Q and believes that she will not Q. However she also has another judgement that it would be better *to* Q and this judgement is not relative to any particular considerations the agent has, it is just an 'unconditional judgement'[8] or evaluative attitude. What distinguishes the weak-willed agent from the continent agent is that the former ends up acting on this judgement despite having a better reason to avoid doing so. On Davidson's view we can still criticise the weak-willed agent because she fails to govern her actions in the appropriate manner; but what is most interesting about this view is that Davidson makes it possible to analyse weak-willed actions from the standpoint of the motivational sceptic about pure practical reason. If this view can satisfactorily account for both weak-willed and ordinary intentional action then the force of the attraction of the PPR view will be greatly diminished.

It might be objected to my project at this stage that the dispute between the PPR and IR views cannot be resolved by figuring out which one can present the most satisfactory account of intentional and weak-willed actions. It has been argued for

---

[8] Davidson uses the word 'judgement' rather loosely to refer to any values, tastes, principles, inclinations and evaluative attitudes generally that agents may have ('Replies to Esays I-IX', *Essays on Davidson: Actions and Events*, p.202).

example, that the IR view is firmly committed to an outmoded eighteenth-century philosophy of mind[9] or that among other faults it relies on an artificial distinction between reason as being purely informational and reason as having normative authority. However I believe that the problem of explaining irrational action *does* indeed go to the heart of the matter. Explaining how it is that we can fail to act as we decide we should act is something which must be accounted for irrespective of whether, for example, having a belief and having a desire turns out to be a unitary state, or whether instrumental reasoning involves at least some minimally substantive conception of an agent's good. These issues are not irrelevant, but the point is that any worthwhile theory of practical reason must begin by including within its purview how we are to account for irrational action (however conceived) as well as intentional action generally. We cannot begin by asking what desires are and how they are related to beliefs. Questions of this nature are circumscribed by what it is a theory of practical reason is supposed to do – explain how we appraise our actions and how it is we govern our reasoning about what to do by norms we endorse or believe to be binding. What we think is happening when the usual process fails to result in action will in itself constrain any account we develop.

Any account of practical reason that too readily sets aside cases of irrational action thereby insulate themselves from precisely the sorts of cases which a theory of practical reason needs to address. Such a theory cannot be a satisfactory position to hold. In other words, any account that begins by arguing that in so far as we are rational, reasons *do* move us to act has illegitimately imposed what we might call a *defeasibility condition* on a theory of practical reason. In this case what such a condition does is to turn what appears to be hard cases that might give us good reason

[9] This is McDowell's (1981) well known criticism of non-cognitivism. See Simon Blackburn (1981) for a response.

to revise our theory of practical reason into merely unusual instances that can be explained away not as counter-examples but as situations where the agent's reasoning capacity is below par, as it were. In particular, weak-willed actions are said to be *irrational*, and of course the presumption is that we only have to consider what is happening when agents reason about what to do *rationally*. Examples of irrational action therefore, we are supposed to believe, cannot circumscribe how we understand practical reasoning generally and in particular the connection between having an all-things-considered reason for action and being motivated to act accordingly.

But before we examine this strategy let us take another look at the internalism requirement and anticipate our discussion of the more straightforward (even if ultimately more problematic) strong internalism version of the internalism requirement, which attempts to resolve the inconsistency in our intuitions about action without arguing for a defeasibility condition. As stated above this requirement really does little more than express our insight about action being reliably connected to an agent's judgement or consideration. It is counterintuitive that agents can act against or fail to act according to their own recognised better reasons for action all-things-considered. So the question then becomes how we are to think of the connection between having a reason for action all-things-considered and being motivated to act. If we argue that this connection is a *necessary* one then we are committed to a *strong* internalism requirement because we will be arguing in effect that if an agent holds an all-things-considered reason for action then necessarily she will act. The problem with this view however is that if we take into account weak-willed action *as described* (namely, as action against one's all-things-considered judgement) then we are forced either to reject this strong version of internalism altogether or to deny that weak-willed actions *as described* actually occur. The latter option is the most favoured one. Despite what some might think it is far from typical to find anyone rejecting the strong internalism

requirement because of worries about weak-willed action.[10] More frequently found is a willingness to reject our understanding of weak-willed action as described, so that, for example, the weak-willed agent does not fail to act on her all-things-considered reasons, rather she never gets as far as correctly forming the proper intention to act based on her all-things-considered reasons.

Many believe that following this strategy they can satisfactorily explain weakness of will without having to resort to the outright denial of the strong internalism requirement. Instead, they favour an explanation that denies that weakness of will really is action against all-things-considered reasons. Weak-willed agents are supposed not to form the type of reason for action that is necessarily connected to motivation. Yet resolving the problem of akrasia in this manner is fraught with difficulties, not least of which involves attempting to show that akratic agents are unable to form the right type of better judgement that *would* be necessarily connected to their motivation. While this provides the best chance for those accepting the strong interpretation of the internalism requirement to resolve the problem of akrasia, it is, as we will see, ultimately an unsatisfactory solution.

The most common solution is to see the connection between having an all-things-considered reason for action and being motivated to act as *defeasible* by an agent's weaknesses, compulsions, irrational desires etc. Agents will typically act on an all-things-considered reason, we are told, and when they do not then we do not have to conclude that the fault lies with their all-things-considered reason. The problem must lie instead with some interfering desire or compulsion they have. However there are marked differences between the various attempts to adopt this strategy. This strategy is favoured by both contemporary non-naturalists (especially Kantians) about practical

[10] David Brink (1986) and (1989) comes closest to denying internalism based on the existence of weak-willed agents.

reason as well as various kinds of humean-inspired naturalists. It is typically held on both these views that if an agent judges that it would be best, all-things-considered, to $\phi$ in circumstance C then either she is motivated to $\phi$ in C or she is *practically irrational*, and we can go on to give numerous arguments concerning what is happening to the practical reasoning of an agent whose behaviour is, by their own lights, against their all-things-considered reason for acting. We can call this weakness of will and this in turn can be described as acting against better judgement, acting on irresistible desires, self-deception etc. The point is that once we allow for the possibility of weak-willed action by categorising it as practically irrational, and making irrational action a defeasibility condition on the internalism requirement, then we have effectively *dissolved* the *problem* of weakness of will since it can no longer be seen as a counter-example or as evidence of a lack of a necessary connection between having an all-things-considered reason to act and being motivated to act accordingly.

Let us call those who put a defeasibility condition on the internalism requirement *weak internalists.* What weak internalists want to avoid is the conclusion that strong internalists come to, namely that the agent who fails to $\phi$ in C never had the judgement that it would be best to $\phi$ in C in the first place. In addition to the fact that this conclusion does not fit the phenomenology of the experience of being *obligated* (since, apparently, we can still feel obligated to act even when we do not seem to *desire* to do so), it prevents us from distinguishing action that follows reliably upon all-things-considered reasons from action that is weak-willed or otherwise contrary to an agent's better judgement. If we must always assume that an agent who fails to do what she says she has an overriding (or all-things-considered) reason to do, must not have had that reason in the first place and must have acted on another reason that was her real all-things-considered reason for action, then it threatens to become impossible to distinguish between the weak-willed agent and the ordinary continent agent. The

former would appear frequently to act contrary to how she might be supposed to act given that she declares herself to have all-things-considered reasons that do not fit the action she in fact performs. But on the strong internalist view this seems more like self-deception than weakness of will. For if the agent just does not know what are the *real* reasons motivating her to act, and she believes herself to have different reasons than those which the strong internalist view would impute to her, then this agent seems amazingly unaware of her own reasoning about what to do. This not only is blatantly counter-intuitive since we think that weak-willed agents are characterised by *knowingly* acting against what *they take to be* their all-things-considered reasons for action, but it is difficult for independent reasons to rest easy with ascribing to agents such a degree of self-deception about their own reasons for action.

These attempts to solve the problem of weakness of will bring to the fore the inadequacies both in the strong internalist's defence of the internalism requirement as a claim asserting a necessary connection between an all-things-considered reason and motivation to act, and the weak internalist's defence of the internalism requirement as calling for defeasibility conditions. It is not sufficient, I will attempt to show, to argue that the state of holding an all-things-considered judgement or reason for action is motivational in virtue of its having the authority of an *endorsed* reason and then to continue by allowing various forms of 'irrationality' to defeat an otherwise rational agent's all-things-considered reason for action. For example, Korsgaard argues that an agent who fails to act on a reason she considers compelling is just showing the limits of her rationality; she just may be irrational but this is not sufficient grounds for scepticism about the internalism requirement as favoured by those opting for the weak reading of the requirement.[11] Korsgaard argues that once we realise that what we are calling weak internalism is defeasible, or that it has room for various forms of

[11] 'Skepticism about Practical Reason', *The Journal of Philosophy,* 1986, pp.24-6.

'irrationality', then scepticism about whether accepting that a moral judgement is sufficient for action loses whatever force it has.

What I suspect has made such a view appear to be so plausible is an influential solution to the difficulties raised when we distinguish between *motivating* and *normative* reasons for action. Some reasons for action are motives but not all motives are reasons for regarding an action as *right* or *obligatory.*[12] For example, John desires to wake the neighbours just to annoy them and to this extent we can say that he has a motivational reason to act (i.e. he simply *wants* to). But if we go on to say that John is therefore justified in waking the neighbours just to annoy them then we are conflating motivation with justification: wanting to act does not mean the action is therefore justifiable. Acknowledging this distinction allows us to carefully avoid making the mistake of thinking that all reasons for action are *motives* for action. Korsgaard is aware that it is not enough to argue that if an action is morally obligatory then there is *sufficient* reason to act. This would be an identification of *reasons* for action with *motives* for action and this is a rather implausible position. Her solution is to argue not that normative reasons must necessarily move us to act but that when normative reasons *do* move us, they move us *necessarily.* That is, *in so far as we are rational* reasons for action will move us to act. We can see that it is a natural progression from this to arguing that the weak internalism requirement is defeasible by various forms of *irrationality.* If normative reasons for action necessarily move us to act only in so far as we are rational then what we have to do is argue that agents are *capable* of acting on their normative reasons, that is, of acting rationally. But according to Korsgaard just because we are capable of acting rationally does not mean that we necessarily and always act according to our normative reasons. Just as being capable of logically

[12] W.D. Falk *Ought, Reasons and Morality* (Ithaca: Cornell University Press, 1986);William Frankena 'Obligation and Motivation in Recent Moral Philosophy' in K.E. Goodpaster ed., *Perspectives on Morality: Essays by William K. Frankena* (London: University of Notre Dame Press, 1976).

drawing a conclusion does not imply that we always do draw logical conclusions, so being capable of rational action does not imply that we will always act rationally; but when we do act according to our best or all-things-considered reason we are motivated to do so necessarily.

However in what follows we will see that in fact weak internalism fails to show that it is impossible even for an agent qua rational to have a normative reason to act without the motivation to act, and it is the attempt to dissolve the problem of weakness of will that brings this to the fore. What will be the results of this subtle interplay between the problem of practical reason and motivation and the problem of properly accounting for the phenomenon of weakness of will? Hopefully it will become clear that the solutions put forward by these two versions of internalism not only fail to account for all types of relevant cases but to a certain extent they also beg the question regarding the nature of the connection between having a reason and being motivated to act. Strong internalists (such as Davidson) can only tell us that akratics suffer from a form of inferential failure or a breakdown in the processes of practical deliberation, but we will see that this solution fails to account for some key features of akratic action. Weak internalism on the other hand, does allow akratic agents to have 'proper' reasons for action (i.e. reasons of the correct *form*), but their practical deliberation about what to do is defeated by something contrary in their dispositional natures. Weakness of will is a *practical* failure rather than an *inferential* failure (in the logic of intentional action) on the weak, as opposed to the strong internalist view. But we shall see that in attempting to account for all the relevant features of akratic action weak internalists are forced to further weaken their view of the motivational efficacy of reasons for action. When they do this however, it becomes clear that these theorists really have no account of akrasia at all beyond what we already know intuitively – that something 'goes wrong' dispositionally that prevents the agent from acting as she thinks best all-things-

considered. But no matter which of the conflicting views we take of weakness of will we can always say that there is *something* wrong with the agent. So the only plausible weak internalist solution effectively just restates the phenomenon to be explained. What we would like to have then is an explanation of akrasia that covers all the relevant cases without either assuming from the outset that the connection between judgement and motivation is a necessary one on the one hand, or simply restating the problem on the other.

What will emerge as we proceed is that the real target of the PPR criticisms ought to be the IR view which says that weak-willed agents never had the correct judgement in the first place (i.e. the 'no genuine judgement view'). However when we combine the 'no genuine judgement' view (warts and all) with our dissatisfaction with motivational internalism in its various existing guises it *is* almost tempting to accept the PPR arguments – especially if we find successful the argument that the IR view cannot provide an account of the instrumental principle that is sufficiently normative to stand alone as the sole principle of practical reason. However I attempt to argue that we are not put in the position of having to abandon an IR view in favour of a PPR one. But I do not intend to pursue an argument to the effect that Kantians fail to establish categorical principles of practical reason. This would be an entire work in its own right. Yet another neo-Kantian strategy troubles me. The idea seems to be that since opposing views on practical reason (i.e. Humean views) cannot establish the instrumental principle as the sole normative principle of practical reason, then an IR view of practical reason is doomed from the outset and it is the examination of weakness of will that is supposed to make this clear to us. The PPR proponents seem to treat this apparent difficulty as something like a reductio (of the IR view) which lends further support to their own view of practical reason.

What I try to argue here is that the PPR view is not the only game in town when it comes to a workable theory of practical reason. I agree with those of a PPR persuasion that weakness of will is important in all of this. However I do not agree with their argument that an IR view is in the implausible position of being unable to adequately distinguish the weak-willed agent from the rational agent. Rather, I argue that an IR view does a better job of accounting for weakness of will since it does not start by assuming that it is a motivational problem only. As indicated above, those holding the PPR view suggest that the agent who fails to act on her reason for action all-things-considered is suffering from a (psychological) breakdown in her rational practical reasoning and action. However this view is only available on the assumption that endorsing a reason for action guarantees that an agent will act accordingly (*ceteris paribus*) insofar as she is rational. I argue that this is a very problematic assumption to make in any theory of practical reason. Weakness of will complicates matters for any view that claims, on the basis of this assumption, that an agent will automatically act (all things being equal) in accordance with her all-things-considered reason for action. What will hopefully become clear is that having an all-things-considered reason for action is not the final point or step in one's practical reasoning before action takes place. Having such a reason underdetermines an agent's intention to act. As we shall see an intention to act is a certain kind of favourable attitude with important features. It makes most sense for agents to have a favourable attitude towards what they believe they have most reason to do. However I remain doubtful as to whether agents can be rationally bound to act on their reasons in any stronger sense than this. There might be many possible ways in which an IR account of practical reason could be expressed. I just outline my suggestion here. However what is an important step in thinking about a theory of practical reason is that an agent's attitudes regarding her reasons for action play a crucial role in bringing about that action. It is only when we see that weakness

of will or generally considered irrational action cannot conveniently be redescribed or sheltered behind a defeasibility condition that we can see that we ought to re-examine the assumption we make about endorsing a reason for action and being motivated to act on it.

* * *

My comments above will not be fully clear until we have dealt with each of the issues raised in detail. However before we begin a brief review of the contents of the chapters might be helpful.

Part I consists of chapters one and two. These chapters discuss motivational internalism, broadly speaking the view that motivation to act is necessarily connected to the holding of a corresponding judgement about how to act. Chapter one discusses motivational internalism generally and sub-divides internalism into Strong and Weak motivational internalism. Chapter two concerns the three different motivational internalist strategies for dealing with weakness of will. Each strategy is discussed and the problems facing these strategies and facing motivational internalism itself are outlined.

Part II consists of chapters three and four and as a whole deals with the neo-Kantian or Pure Practical Reason (PPR) view mentioned above. On this neo-Kantian view weakness of will and other irrational actions are simply examples of conditions that can cause us to act irrationally, but such actions can no longer be used to deny that there must be principles of conduct other than hypothetical imperatives based on desires. In fact it is argued that weakness of will cannot even be explained on the opposing view of reason as instrumental. Chapter three will examine the considerations that lead PPR theorists to hold such a view before we go on to chapter

Four, where I will contend that those favouring the Pure Practical Reason view end up begging the question regarding weakness of will because of their view that one endorses a reason for action only when one evaluatively concludes that that reason is best, all-things-considered.

Part III attempts, in outline, to tell a somewhat different story about the relation between endorsing a reason and acting on it. The PPR assumption above is one I suggest we should reject and replace with a different understanding of what it is to endorse a reason as one upon which the agent should act. What is disputed here is the PPR view that endorsement amounts to evaluative judgement. I suggest instead that attitudes *do* play a crucial role in endorsement. By arguing that we can explain cases of weakness of will more successfully in terms of attitudes and their stability we show that the PPR criticisms of properly developed IR views are unfounded. Thus the way lies open to further develop this attitudinal account of commitment to action into a full-blown theory of practical reason.

# Part I

## Connecting Judgement and Motivation:

## Motivational Internalism and the Problem of Weakness of Will

The purpose of the chapters in Part I is to show that the problem of weakness of will or akrasia has traditionally been, and still remains, a thorn in the side of those wishing to argue that motivation to act is necessarily connected to the holding of a corresponding judgement about how best to act. This view is more commonly known as *motivational* internalism. Historically, the most popular strategy for solving the problem of weakness of will has been to show that weak-willed action is somehow not the result of the 'proper' processes of practical reasoning and is, instead, the result of rationally defective reasoning about what is best to do. In what follows, however, I show that this strategy can be sub-divided into three main approaches to the problem of weakness of will. Two of these are favoured by what I characterise as *strong* motivational internalism about the connection between judgement and motivation, and one is adopted by the opposing view – *weak* motivational internalism. I argue that of the strong internalist solutions, one is straightforwardly implausible, and despite some success in explaining certain akratic actions, the other solution is ultimately unworkable because of one key assumption that on closer examination turns out to be false. It is only the third solution, offered by weak internalism, that provides the best prospects for an explanation of the phenomenon and the problems it causes. The problem is, however, that when we sufficiently weaken internalism to account for weakness of will what we end up with is a view that fails to satisfactorily account for the *reliability* of the connection between having a judgement or reason for action and being motivated to act; and on this view akratic actions can neither be explained as breakdowns in this

normally reliable connection nor cited as evidence of a lack of such a connection. All we can gain from this position is the view that weakness of will is simply the result of problems with an agent's dispositional character. But we already know that we can give this type of exculpatory explanation for akratic actions; the problem is that this position fails to address the prior question as to whether explanations of this kind are themselves sufficient. We know that weakness of will involves what we might call practical failures, but the key question is whether it also essentially involves something more like a logical failure in the sense that an akratic agent fails to recognise key aspects of the logically necessary features of what it is to have a reason to act at all. What I argue in the following two chapters is that the problem of weakness of will cannot be discussed, never mind solved, independently of the question regarding how to characterise the reliability of the connection between judgement and motivation to act.

# Chapter One

## Weakness of Will and Motivational Internalism (1): the Problems

### Introduction

The problem of weakness or will or *akrasia* has undergone an alteration of status since it was first discussed by Plato in the *Protagoras* and *Republic*. In these early texts, akrasia is a significant problem for a theory of practical reason. If an agent *knows* what will further the good how could she possibly do anything else? Deliberation about what to do only makes sense for Plato in the context of the good and consequently an agent's reasons for action are determined by the good since this is what all rational action aims at. So any agent who deliberates about what to do but fails to do it is a cause for serious concern. Plato ends up by saying that such an agent must have a disordered soul and must not, in fact, be capable of proper deliberation. But this is not just a quick-fix solution to the problem of how it is that an agent can perform an action that she knows to be overruled by her better judgement. Plato simply cannot see how a properly deliberating agent with a correctly ordered soul could possibly act akratically and so he rules out the very possibility of this occurring.[1]

Contemporary interest in akrasia, however, does not share Plato's view of its significance. In fact it would not be incorrect to say that akrasia today has become a much more minor philosophical problem, one simply to be ironed out, as it were. With the publication in 1963 and 1970 respectively of Donald Davidson's highly influential papers, 'Actions, Reasons and Causes' and 'How is weakness of the will possible?'[2] there was a resurgence of interest not only in akrasia but in a causal theory of action

[1] See footnote 2, 'Introduction', p.8.
[2] 'Actions, Reasons and Causes', *Journal of Philosophy* 60 (1963); 'How is Weakness of the Will Possible?', Joel Feinberg (ed.), *Moral Concepts* (Oxford, 1970). Both repeated in Davidson *Essays on Actions and Events* (Oxford, 1980).

more generally[3] and for a brief period that seems to have peaked in the 1980s, akrasia was once again a significant problem not just of moral theory and practical reasoning but of action theory as well.[4] One of the problems that frequently arises in explaining intentional action is that it is commonly thought that any causal theory of intentional action presupposes a necessary connection between an agent's *judgement* of what it is best to do (all things considered) and her *motivation* to act. However this causes great difficulty when we try to explain weak-willed action. For if we believe that akrasia is a real phenomenon (i.e. that it really is action against better judgement) then we may find ourselves facing a dilemma not unlike the one which troubled Plato. On the face of it, if there is a necessary connection between judgement and motivation then there really is no possibility of action occurring that is against one's better judgement. But if weak-willed action does indeed occur then how can we reconcile this fact with the necessary connection between judgement and motivation?

The only way out of this apparent dilemma is either to deny the necessary connection between judgement and motivation or to deny that weak-willed action *really is* action against an agent's better judgement. But Davidson, for one, does not believe that the possibility of akrasia poses any real dilemma for his theory of action, and he argues that we are not, therefore, forced to deny that akrasia really occurs. However, much of the writing that this view inspired has tended to question Davidson's solution to the problem of akrasia. This interest in Davidson's position has more or less waned and it appears that once again akrasia is in the process of being relegated to the minor league of philosophical problems. However I believe this relegation betrays

---

[3] The attempt to naturalise reasons for action by reviving a humean-inspired view that all action is motivated by prior desire was undergoing resurgence of interest around this time. Davidson attempted to "...defend the ancient – and commonsense – position that rationalization is a species of causal explanation." (1980, p.3) This was a response to opposing views held by, for example, Gilbert Ryle (1949); G.E.M. Anscombe (1959); A.I. Melden (1961).

[4] For example Robert Dunn (1987); Justin Gosling (1990); Frank Jackson (1984); Michael Kubara (1975); David Pears (1982,1984); David Pugmire (1982); CCW Taylor (1980).

a fundamental misunderstanding of the real significance of the problem.

What lies behind the failure to see akrasia as a significant problem is a pervasive view concerning the nature of the connection between having a judgement or reason for action and acting upon it. This view, as we have seen, is known as *motivational internalism,* and although there are multiple versions of this view, a reasonable internalist account of rational agency claims that reasons are intrinsically motivating. That is, if S has reason R to do A, then S is thereby motivated to do A, and (ceteris paribus) will do A upon concluding R. The central tension is that akrasia looks to be a case where one's best reason fails to be action-motivating, and so is contrary to internalism. However given the widespread commitment to internalism we can easily understand how akrasia can seem like a problem whose resolution is to be found in its redescription. If weakness of will really is action against better judgement then in order to reconcile it with internalism we have three options: we could (1) deny motivational internalism altogether; (2) deny that weakness of will really is intentional action against better judgement (for example, by arguing that it is essentially about suffering from irresistible desires or some sort of compulsion); or (3) stand firm on internalism and argue instead that weak-willed agents do not form the type of judgement that is necessarily connected to their motivation. Options (2) and (3) together provide the structure of the overwhelming majority of work done on akrasia to date, but to the best of my knowledge no one has ever attempted either to use akrasia to deny internalism, or at least to question the manner in which internalists find it relatively easy to dissolve the problem of akrasia altogether.

In what follows what we shall be examining closely is the way in which internalists find akrasia and all other kinds of practical irrationality to be fairly straightforward problems. It appears to be the current view that so long as motivational internalism allows some defeasibility conditions, practical irrationality can be

acknowledged, categorised, resolved and put away. But *this* is precisely what needs closer examination.

For our purposes we can distinguish between two strains of motivational internalism – *weak* internalism and *strong* internalism. Loosely put, for strong internalism it is a necessary condition of its being the case that an agent has good reason or acknowledges an obligation to X that that agent would, under appropriate conditions, be motivated to X. Weak internalism claims, on the other hand, that sufficient motivation for action is not a necessary condition of holding a reason for action but that *some* degree of motivation to act is necessary. Again there are numerous variations on a theme here, but generically speaking strong internalism has been around for a long time (since Plato) whereas weak internalism is relatively new on the historical stage. What I hope will be clear from what follows is that it is *strong* internalism that has sustained interest in weakness of will since it is this version of internalism which most obviously appears to be inconsistent with the existence of akrasia. While there are some difficulties reconciling the existence of weakness of will with this version of internalism, it is not in the least surprising that this appears to be a rather tough task.

However many strong internalists believe that they have satisfactorily explained weakness of will without having to resort to the outright denial of an internal connection between having a reason and being motivated to act on it. Instead, they favour an explanation that is an amalgam of options (2) and (3) above, that is, denying that akrasia really is action against better judgement and at the same time standing firm on internalism, and arguing that weak-willed agents do not form the type of judgement that is necessarily connected to motivation. Yet resolving the problem of akrasia in this manner is fraught with difficulties, not least of which involves attempting to show that weak-willed agents are unable to form the right type of better judgement that *would* be

necessarily connected to their motivation. While this provides the best chance for strong internalists to resolve the problem of akrasia, it is, I think, ultimately an unsatisfactory solution.

For weak internalists, on the other hand, the solution seems relatively clear. People behave in all sorts of ways such that, if questioned afterwards, they would claim that they acted against their better judgement of what to do. Such actions are practically irrational, since it is an action that is contrary to what the agent believes she ought to do. It might appear to be a consequence here that weak internalists are put in the position of either having to deny akrasia altogether or accept that akrasia gives them no other option but to abandon their claim that a necessary condition of holding a judgement about how best to act is that the agent is motivated to act on it. The problem with holding that there is such a necessary connection between believing that an action is the best thing to do and being motivated to act on it is, as we have seen, that it appears to be inconsistent with the existence of weak-willed action. However weak internalists, aware of the implications that would ensue if this inconsistency was indeed found to obtain, have managed, they think, to make the existence of akrasia consistent with the necessary (but not sufficient) connection between judgement and motivation.

The best way for an internalist to do this is to argue that the necessary connection is *defeasible*. So, if an agent judges that it would be best, all-things-considered, to $\phi$ in circumstance C then either she is motivated to $\phi$ in C or she is *practically irrational*, and the theorist can then go on to offer various characterisations of what is happening to the practical reasoning of an agent whose behaviour is, by their own lights, irrational. We can call this akrasia and this in turn can be described as acting against one's better judgement, acting on irresistible desires, a form of self-deception etc. The point is that once we allow for the possibility of weak-willed action by categorising it as practically irrational, and making irrational action a defeasibility

condition on motivational internalism, then we have effectively *dissolved* the *problem* of akrasia since it can no longer be seen as a counter-example or as evidence of a lack of a necessary connection. What weak internalists want to avoid is the conclusion that strong internalists come to, namely, that the agent who fails to $\phi$ in C never really had the judgement that it would be best to $\phi$ in C in the first place. Despite the fact that this latter view just does not fit the phenomenology of the experience of being *obligated* (since we can still feel obligated or acknowledge that we have good reason to act even when we do not seem to *desire* to do so), it prevents us from distinguishing action that follows upon judgement from action that is weak-willed or otherwise contrary to an agent's better judgement.

However, weak internalism's apparent dissolution of the problem of akrasia is, I will attempt to argue, also an unsatisfactory solution. What weak internalists fail to see is the extent to which the problem of akrasia might require internalism to be weakened to such an extent that it becomes merely a psychological and/or causal (i.e. contingent) claim. As a thesis about our dispositional character weak internalism is not an unattractive position, and even many externalists could find its core claims acceptable. However if the claim is that it is a *necessity* that making or assenting to a judgement necessitates at least *some* inclination to act, then even weak internalists have not done enough to prove their claim. Their attempt to dissolve the problem of akrasia will bring to the fore the inadequacies both in the strong internalists' defence of internalism and in the weak internalists' defence of internalism as permitting defeasibility conditions.

It is not sufficient, as I will attempt to show, to argue that the state of holding an all-things-considered judgement is identical with, or necessarily connected to, the state of being motivated in a certain way, and then to continue by allowing for various forms of 'irrationality' as a defeasibility condition. For example, as we briefly introduced earlier, Christine Korsgaard argues that an agent who fails to act on a reason she

considers compelling is just showing the limits of her rationality; she might well indeed be irrational but this is not sufficient grounds for scepticism about what we are calling weak internalism.[5] Korsgaard argues that once we realise that weak internalism is defeasible, or that it has room for various forms of 'irrationality', then scepticism about whether accepting that a moral or evaluative judgement is sufficient for action loses whatever force it has. Korsgaard is not alone here but I suggested earlier that perhaps what has made such a view appear to be so plausible is an influential solution to the difficulties raised when we distinguish between *motivating* and *normative* reasons for action. Some reasons for action are motives but not all motives are reasons for regarding an action as *right* or *obligatory.*[6] This distinction allows us to carefully avoid making the mistake of thinking that all reasons for action are *motives* for action. Korsgaard is aware that it is not enough to argue that if an action is morally obligatory then there is sufficient reason to act. This would be an identification of *reasons* for action with *motives* for action and this is a rather implausible position. The solution here is to argue not that normative reasons must necessarily move us to act but that when normative reasons do move us, they move us necessarily. So, *in so far as we are rational* (normative) reasons for action will move us to act. We can see that it is a natural progression from this to arguing that weak internalism is defeasible by various forms of *irrationality*. If normative reasons for action necessarily move us to act only in so far as we are rational then what we have to do is argue that agents are *capable* of acting on their normative reasons or, of acting rationally. But just because we are capable of acting rationally does not mean that we necessarily always act according to our normative reasons all the time. Just as being capable of logically drawing a conclusion does not imply that we always do draw logical conclusions, so being

---

[5] 'Skepticism about Practical Reason', *The Journal of Philosophy,* 1986, pp.24-6.

[6] W.D. Falk *Ought, Reasons and Morality* (Ithaca: Cornell University Press, 1986); William Frankena 'Obligation and Motivation in Recent Moral Philosophy' in K.E. Goodpaster ed., *Perspectives on Morality: Essays by William K. Frankena* (London: University of Notre Dame Press, 1976).

capable of rational action does not imply that we will always act rationally. However, in what follows we will see that in fact weak internalism fails to show that it is impossible even for an agent *qua rational* to have a normative reason to act without the motivation to act and it is precisely this attempt to dissolve the problem of akrasia that brings this to the fore.

What we will see is that while weak internalism works better as an *a posteriori* position rather than as an *a priori* one, ultimately the concessions it has to make in order to account for different types of cases of weak-willed action make it an unattractive position. So what are the implications for weakness of will? Hopefully it will become clear that the solutions put forward by these two types of motivational internalism not only fail to account for all types of cases but to a certain extent they also beg the question regarding the nature of the connection between having a reason and being motivated to act. Strong internalism can only tell us that weak-willed agents suffer from a form of inferential failure or a breakdown in the processes of practical deliberation, but we will see that this solution fails to account for some key features of weak-willed action. Weak internalism does allow weak-willed agents to have 'proper' reasons for action (i.e. reasons of the correct *form*) and so their practical deliberation about what to do is defeated by something contrary in their dispositional natures. Akrasia then is a practical failure rather than an inferential failure on this view. But weak internalism also cannot account for all the relevant features of weak-willed action and in order to do so it has to further weaken its view of the motivational efficacy of reasons for action. When we do this, however, it becomes clear that we really have no explanation for akrasia at all beyond what we already know – that something 'goes wrong' dispositionally that prevents the agent from acting as she thinks best all-things-considered. But no matter what view we take of akrasia we can always say that there is something wrong dispositionally. Even Davidson, the advocate of the inferential

failure view, admits as much, so the only plausible weak internalist solution effectively just restates the phenomenon to be explained. What we need to do then is find an explanation of akrasia that covers all possible cases without assuming from the outset that the connection between judgement and motivation is a *necessary* one on the one hand, and without simply restating the problem on the other.

## 1.1 Weakness of will and the explanation of intentional action: a dilemma

As we have seen the problem of akrasia appears to generate a great deal of worry due to the common perception that it casts doubt over the old, but still controversial view on the explanation of intentional action, the causal theory of action. With the resurgence from the 1950's of interest in explaining action as caused by an agent's beliefs and desires, akrasia has come to be understood as action against one's judgement of what it is best to do *all-things-considered*. Of course there are those who disagree about this and prefer to talk about akrasia in terms of the irresistibleness of desire or in terms of compelled action.[7] But despite the wide range of views on what exactly constitutes a case of akrasia, the majority of thinkers on the subject now agree that the phenomenon to be explained is how it is that we believe that agents can act freely against their judgement of what is best to do.[8] But whether we accept that this is indeed what

[7] David Pugmire, 'Motivated Irrationality Part I', *PAS* 56 (1982); Alfred Mele *Irrationality* (Oxford, 1987), 'Is Weak-willed Action Unfree?' *Philosophical and Phenomenological Research* 46 (1986); Gary Watson 'Skepticism about Weakness of Will' *Philosophical Review* 86 (1977); Michael Kubara 'Acrasia, Human Agency and Normative Psychology' *Canadian Journal of Philosophy* 5 (1975).

[8] The issue of whether or not the weak-willed agent acts intentionally is dealt with in the final chapter of this thesis. But very briefly, the dispute concerns the extent to which an agent who fails to act on her self-addressed best judgement is acting intentionally. If we argue that intentional action is always action according to one's best judgement the difficulty is that it makes *unintentional* actions done for reasons other than that the proposed action is *the best*. The problem is, however, that it does seem that many actions are performed for reasons other than that they are simply best, say for example, a choice between equally desirable alternatives. If we adopt this criteria for intentional actions then not only do we exclude weak-willed actions from being intentional but also many typical and unproblematic actions also. With

happens when someone acts akratically or we think that acting against better judgement is more accurately described as a situation where desire has overcome our self-control, we still share a common problem.

To put this problem crudely, it is argued that when an agent performs an action the standard explanation is that she did so because she had a reason to act, and so we say that her reason was *the cause* of her action. We generally take this a step further and say, too, that the action is an intentional action because the agent had a reason for acting and that reason was the cause of her action. The problems start here. When we say that an agent had a reason to act as she did we generally mean that she thought this was her strongest reason for action, or, *the best thing to do* after consideration of the options available. Any agent acting against such a this judgement can be accused of acting on a lesser reason – one which she has already judged to be overruled by the reason she considered best. For example, Jane is faced with a choice of whether to meet her friend for coffee now or later this evening. Jane thinks about her day and knows that if she goes now she will break her concentration and may even end up taking a longer coffee break than she desires or (in this case) believes she deserves. Since Jane is diligent (today at least), she thinks that a coffee break with her friend would be more appealing in the evening because she will be finished with her day's work and as a result she will feel good about things. So Jane decides to meet her friend later and sits back down at her desk. When we talk about *why* Jane sits back down at her desk we would say that she does so because she has decided, for a variety of

---

this dispute in mind all the following accept that akrasia is action against one's best judgement, but action done for reasons as opposed to being a matter of compulsion. Frank Jackson 'Weakness of Will', *Mind* 1984, Michael Stocker, 'Akrasia: The Unity of the Good, Commensurability, and Comparability' from *Plural and Conflicting Values* (Oxford: Clarendon, 1990); Davidson 'How is Weakness of the Will Possible?' *Essays on Actions and Events* (Oxford, 1980); Michael Bratman's 'Practical Reasoning and Weakness of the Will' (*Noûs*, 1979); David Pears 'How Easy is Akrasia?' *Philosophia* 11 (1982), *Motivated Irrationality* (Oxford: Clarendon, 1984); Alfred Mele *Irrationality* (New York: Oxford UP, 1987); Elizabeth Rapaport 'Explaining Moral Weakness' *Philosophical Studies* 24 (1973); A.O. Rorty 'Where does the Weak-willed Break take Place?' *Australasian Journal of Philosophy* 58 (1980).

reasons, to go for coffee later on instead of right now. This is Jane's *reason* for her action and it is the *cause* of her action. We even go so far as to say that Jane is acting on her *strongest* reason and it would not be incorrect even to say that Jane *really wants* to stay and get her work done.

But what if Jane, having decided to go for coffee later, gets up and despite her decision goes for coffee now? This is the kind of example that is often cited when we talk about weakness of will and what is perplexing is that if we are to explain actions in terms of the strongest reasons agents have for acting then how can we explain Jane's weak-willed action? By her own admission what she has decided is that it is better to work now and go for coffee later than to go for coffee now. But when Jane goes for coffee *now* we are quite puzzled about her behaviour. How can we explain it? Well, we could say that she was being insincere in the first place. Maybe she was trying to make a good impression on a colleague by appearing to be more concerned about her work than coffee breaks. Or maybe she is sincere but simply deceives herself – perhaps she fails to see that *wanting* to be more concerned about one's work does not mean that one *will* be more concerned. But then again we could try to argue that Jane is just overcome by her desire for coffee and chat despite her good intentions.

But are any of these possible explanations sufficient? I think not. The reason why none of these explanations are particularly illuminating is that we would have to conclude either that Jane somehow does not really think that going for coffee later is the best thing to do or that her action is rendered unfree or unintentional. If she is insincere then she probably believes that appearing diligent to one's colleagues is the best thing to do and *not* that going for coffee later is best. If she is a victim of self-deception then it is likely that she thinks that wanting to be more concerned about work will carry her through the temptations of the day, but of course this too, does not amount to the belief that going for coffee later is the best thing to do. And lastly, if she

is overcome by her desire then we may legitimately wonder whether her subsequent action can be called free and intentional at all.

So, the reason why none of these explanations is sufficient in this case is that they either overlook what is plainly evident – that Jane thinks that going for coffee later is best – or they seem to distort the phenomenon of akrasia by redescribing it as unfree. Insincerity and self-deception do not give us any explanation for why she apparently acts against this judgement *while she holds it*. If she deceives herself or is insincere then she does not hold this judgement at all and so we get no satisfaction here. Overwhelming desire actually has a somewhat stronger explanatory power because we could try to argue that the reason why Jane acts against her belief about what is best to do is that she is simply overcome by desire. She still has the same judgement about what to do but she just cannot do it – being swept along powerlessly is the best analogy here.

However before we get too excited by the possibility of a solution along these lines it has to be said that overwhelming desire as an explanation of akrasia also appears to be unsatisfactory. The reason for this is that agents like Jane are not just swept along powerlessly. Jane knows what she is doing – she knows that going for coffee later is the best thing to do and she knows that she is doing what she believes not to be best. Saying that Jane is overcome by desire means that she 'acts' in such a way as to find herself standing up from her desk and moving towards the door, lifting the coffee cup etc. If this is what we think happens then if we continue and argue that Jane's resulting action is also intentional the only role left for intention is some sort of awareness on the agent's part of what movements she is making, not unlike the awareness of a bystander witnessing certain events. Now some readers may want to argue that we are making a mistake here because we are illegitimately conflating

intentional action with free action.[9] However when an agent forms an intention to act we commonly believe that she will, all things being equal, act accordingly. If Jane has formed an intention to go for coffee later but does not do so, well then we may think that she *subsequently* had an overwhelming desire and acts *contrary* to her intention. The crucial question that surfaces here concerns the role we think intention plays in bringing about the agent's action.

The role of desire in weak-willed action requires a more sustained treatment than my few comments here have provided, but there are some interesting issues emerging that require more immediate attention. The first of these concerns what a causal theory of action is actually trying *to do*. In essence it attempts to give a satisfactory belief-desire account of action. That much is generally clear, but what is often less than clear is how such an account is to be constructed. Just how exactly are beliefs and desires supposed to come together so as to result in an action? The best way to clarify this is to look a little more closely at what a causal theory is supposed to accomplish. Now the common response here is that the beliefs and desires an agent holds act as the agent's reasons for (or, together they 'rationalise') her action, and we say that the reason for which the action was performed is the *cause* of the action. While this is correct, there is a danger that it might obscure what we are really saying here – that action follows upon judgement. So when an agent judges (in the broadest sense) that she should X (all-things-considered) then, all things being equal, she will X. Clearly then, there is an important connection between judgement and action but it is not clear what the status of this connection is. Does judgement *reliably produce* action or does judgement *conceptually entail* that if an agent judges (or intends, decides, etc) to X at t and is physically and psychologically able then the agent X's at t? The fear is that the capturing the connection in the former way leaves us with an unsatisfactory

[9] Gary Watson (1977) p.334.

degree of contingency. Are we happy to see practical reason founded upon a connection that could have been otherwise? However if we claim instead that judgement is conceptually linked with action then the danger is that we might end up with a connection that is worryingly strong.

Here we can see that the problem of akrasia arises right from the beginning of any account of intentional action that attempts to explain action in terms of agents' reasons. The apparent dilemma arises when we want to hold both that there is a necessary connection between an agent's reason for acting and her intentional action and that weak-willed action really is action against better judgement. But if there is such a *necessary* connection between better judgement and action then it appears to be very difficult to allow for the possibility of the agent making an all-things-considered judgement and *not* acting accordingly. If Jane judges that going for coffee later is best and this judgement is *necessarily* connected to action then Jane will sit back down to work and it will not be possible for her to intend to act contrary to her better judgement. This necessary connection serves to copper-fasten the link between judgement and action and threatens to make any kind of *intentional* action against one's best judgement impossible. With judgement necessary for action in this sense Jane can rest assured that she will never perform an uncompelled action that is contrary to her best judgement. She may well experience overwhelming desire and give in occasionally but any such action could not properly be called weak-willed, rather it is compelled and unfree action or in some different way non-rational. The only way to make akrasia and this necessary connection consistent is to redescribe akrasia as compelled or somehow unfree action.

If we are determined to hold a motivational internalist view of the production of intentional action then two options for explaining akrasia face us: (A) to deny that akrasia is genuinely action against better judgement at all, or (B) to attempt to

redescribe the type of judgement the agent forms in apparently weak-willed situations. To pursue the latter redescription we would have to argue that it is the agent's *judgement* that is somehow not genuine because, for the strong internalist at least, evaluative judgements *properly formed* are necessarily connected to action. There is no possible way any agent with a proper evaluative judgement might *not* act accordingly.

Plato is perhaps most notable for favouring the strong internalist position with his initial claim in the *Protagoras* that no one does wrong willingly as well as his later view in the *Republic* that akrasia is not uncompelled free action but a disruption to the order of control in one's soul. R.M. Hare actually argues for both these explanations of akrasia and argues that agents in situations such as Jane's above fail to assent to or recognise the prescriptivity of the judgements they make, hence they do not make proper evaluative judgements at all.[10] Donald Davidson prefers the second strong internalist option, arguing, as we will see in a moment, that while akrasia is indeed a genuine phenomenon, weak-willed agents seem incapable of forming the type of judgement or intention to act that non-weak-willed agents make – one that is immune from akrasia.

What both weak internalists and strong internalists have in common here is that they all believe that akrasia does not place them in the dilemma we described above and that it can be made consistent with internalism. Davidson, for example, argues that the existence of akrasia does not pose any threat to his causal theory of action and so we are not forced to deny the possibility of weak-willed action. What we are forced to do however, is to redescribe it in light of the way in which he believes practical reason results in action. Weakness of will for Davidson concerns incontinence of *belief* or *judgement,* as it were, but not incontinence of action. A weak-willed agent's *action* is always performed intentionally because it is always performed with the evaluation or

[10] R.M. Hare, 'Backsliding', *Freedom and Reason* (Oxford University Press, 1963).

judgement that the action is, in some sense, the best. Davidson suggests that the real problem in these situations is not the agent's action but her belief or intention *formation*.

Briefly what happens is this. The agent judges that *all things considered* it would be better *not* to Q and believes that she will not Q. However she also has another judgement that it would be best to Q and this judgement is not relative to any particular considerations the agent has, it is just an unconditional judgement or evaluative attitude.[11]

What distinguishes that weak-willed agent from the continent agent is that the former ends up acting on this latter judgement despite having a better reason all-things-considered to avoid doing so. *Why* the weak-willed does this is not of as much concern to Davidson as the form her practical reasoning takes in these situations; and so his version of weakness of will is a qualified description of the traditional (strong internalist) account which sees weak-willed action as simply action against better judgement all-things-considered. It *is* indeed action against better judgement all-things-considered but Davidson does not agree that such action is impossible or inconsistent with an internalist conception of intentional action. There is no dilemma as far as Davidson is concerned. Quite simply, it just is possible to act on a judgement that is contrary to one's best judgement all-things-considered even though there is a necessary connection between judgement and action. Agents act intentionally, but they do not form the *right* intention – one based on their best judgement all-things-considered – rather the intention they do form is based on a reason they have already considered and overruled. Why anyone would do this is another issue, and Davidson is not in the business of offering a psychological explanation of what brings about

[11]As we noted in the introduction, Davidson uses the word 'judgement' rather loosely to refer to any values, tastes, principles, inclinations and evaluative attitudes generally that agents may have. (Davidson, (1985), p.202)

weakness of will in some agents. It could be self-deception, overwhelming desires or anything else but whatever the cause, *this*, Davidson argues, is the form of the reasoning. [12]

This attempt to make akrasia consistent with an internalist conception of intentional action has been extremely influential, so much so that akrasia has been considered not to pose a serious threat to any form of internalism in much of the subsequent writing on the topic. But this type of solution to the problem of akrasia is not exactly a new one. Aristotle, and to a lesser extent Plato, proposed that since evaluative judgements are action-guiding, weak-willed agents must not be forming their action-guiding judgements in the right way. There is something disordered about a weak-willed agent's soul, according to Plato, that makes them form action-guiding judgements in the wrong way; and according to Aristotle, such agents are suffering from a form of "forgetting" about their better judgement, so they end up acting on a lesser reason because they have "forgotten" their better reason.[13] This type of argument has resulted in akrasia not being widely considered as placing internalists in the dilemma we outlined above, and yet it does not require that they deny that akrasia is action against one's *better* judgement.

However it is frequently denied that weak-willed action really is action against better judgement in the sense Davidson outlines. On this view, akrasia does not present internalists with a dilemma because akrasia just does *not* involve acting on a lesser reason for action, one the agent has already considered to be outweighed by her stronger reason that she thinks she should, all-things-considered, act on. Akrasia is more accurately explained as the action that results when an agent acts in accordance with her wants and desires that have *not* evolved according to the dictates of her reason. Weak-willed action is action that is not in accord with an agent's reason. So we cannot

[12] See Davidson (1970a), p.42.
[13] Plato, *Protagoras*351-8, *Republic* 436-9, Aristotle *Nichomachean Ethics* bk. VII.

classify weak-willed action as acting on a lesser reason for action. In fact such action is characterised as being completely outside the ordinary process of practical reasoning. A typical version of this argument is put forward by David Pears: "passion pushes reason from its throne…The idea is that your thoughts ought to control your actions and akrasia occurs when they do not control them".[14]

A great deal of this sort of argument is the result of a rejection of one or both of two principles of practical reason that Davidson puts forward and defends.[15] He proposes that it is true to claim that (P1) if an agent wants to do X more than she wants to do *Y* (and she can do either) then she will do X (provided she does either); and (P2) if an agent judges that it would be better for her to do X rather than *Y* then she will want to do X more than *Y*.

Anyone who accepts this second principle does so on the basis that there is some sort of necessary connection between *judging* and *wanting*. Now everyone will agree that, ceteris paribus, if an agent wants to do something then she will act. Stating this does not commit us to any particular claim about the nature of the connection between wanting and acting. This is as acceptable to externalists who argue inductively that this is a *reliable* connection as it is to internalists who argue that there is a *necessary* connection here. However the reason why Davidson's solution to the problem of akrasia has provoked more debate than agreement is that the second claim connecting judging and wanting, hence judging and acting, in effect puts unnecessary restrictions on the form intentions are to take, and this has wider ramifications beyond weakness of will. The chief complaints fall into two main categories: (1) that acceptance of the connection between judgement and action is based on a confusion or failure to make a relevant distinction. For example, Elizabeth Rapaport argues that

---

[14] 'How Easy is Akrasia?', *Philosophia,* 1982, pp.1-2, p.33. See also Frank Jackson, 'Weakness of Will', *Mind,* 1984 for a similar view.

[15] These are two principles which he no doubt gets from Hume. See *A Treatise of Human Nature*, 410.

Davidson "fails to distinguish between wants in the sense of particular desires and wants in the sense of rational preferences or over-all self-interest."[16] The objection here is that Davidson just fails to recognise that an agent's strongest reason for action is not always the *cause* of the action. A satisfactory account of the role of desire in action production could better solve many of Davidson's problems. David Pears, Gary Watson, C.C.W. Taylor and Robert Richman make similar arguments.[17] And (2) that Davidson's characterisation of the necessity of the internal connection between judgement and motivation is too strong.

There are two principal strategies which have been adopted to argue that these objections are justified. For (1) above the strategy is to argue that Davidson does not fully understand the concept of *desire*. It is a mistake, so the objection goes, to think that the connection between judgement and motivation must be about *wanting* or *desiring* what one judges better. This connection would be better captured by using an evaluative term such as 'ranking'. What agents are doing when they judge X better than Y is ranking or *valuing* the doing of X higher than the doing of Y. The next move is to argue that agents can and often do value certain courses of action contrary to their particular desires and wants. By making this distinction between particular wants or desires and ranking or valuing we can allow for the existence of weak-willed actions without having to rely on a simplistic conception of the role of desires in practical reasoning.[18]

The strategy for (2) above is to identify a *true* principle which resembles the Davidsonian connection between judgement and action in an attempt to explain why a false principle such as the one proposed by Davidson might seem true to some

---

[16] 'Explaining Moral Weakness', *Philosophical Studies* 24 (1973) pp.180-1.
[17] Rapaport ibid.; Pears *Motivated Irrationality* (Oxford: Clarendon, 1984); Watson 'Skepticism about Weakness of Will' *Philosophical Review* 86 (1977) pp.320-1; Taylor 'Plato, Hare and Davidson on *Akrasia'* Mind 89 (1980); Richman 'Acrasia and Practical Reasoning' *Pacific Philosophical Quarterly* 61 (1980); Alfred Mele *Irrationality* (NY: Oxford UP, 1987).
[18] Rapaport, ibid.

philosophers. Robert Audi, for example, argues that Davidson's connection does indeed grasp a truth – that there is an internal connection between evaluation and motivation – but Davidson misconstrues the nature of the connection by too tightly connecting evaluation and motivation. The denial of this connection as it stands is made even more plausible by attempting to explain how a motivational imbalance can occur. It frequently happens that what we believe to be good or valuable can fail to attract us and what we know to be bad or unworthy of valuing often does attract us.[19]

However with the widespread adoption of these strategies what has happened is that solving the problem of akrasia is considered to be a matter of showing that once it is understood that the connection between wanting, judging and acting is not as tight a connection as Davidson (and indeed Plato and Hare) thinks it to be, we can see that we can avoid getting ourselves impaled on the horns of the dilemma Davidson unwittingly ends up in. Some tweaking with the Davidsonian solution is thus most frequently favoured.

The focus of those in dispute with Davidson tends to be on the issue of the form of the inferential failure that characterises weak-willed practical 'reasoning' and on the problem raised by the claim that an intention to act is always based on the agent's *best* judgement all-things-considered. If intentions to act are one's best judgements and weak-willed actions are performed because an agent intends according to a *lesser* judgement then how do we account for the many actions that are performed when the agent does not believe she is doing what is *best* at the time? And what happens when agents face equally desirable alternatives – the problem of Buridan's Ass? It seems that Davidson's solution to the akrasia problem has the consequence of making many other, non-weak-willed actions weak-willed or otherwise unexplainable. These are real

[19] See E.g. Robert Dunn (1987), Michael Stocker (1979), Alfred Mele (1987), A. O. Rorty (1980) and Robert Audi (1990).

problems to be answered satisfactorily by any account of intentional action, and we will be returning to these issues in Chapter Five. For now, however, I believe we should examine some habitually overshadowed further issues about the nature of this internal connection between judgement and motivation that go to the heart of the problem of akrasia. All too often we find that the nature of this connection is assumed to be a necessary one and so the discussion of akrasia tends to take place in another arena. For example we are told that it must be a problem about strength of desire, that a necessary connection between judgement and motivation or action is not inconsistent with the possibility of overwhelming desire, and that our ignorance of the nature of desire is something which has been largely ignored. This type of argument is part of a search for dispositional features that can break our normally reliable practical reasoning.

There are many proffered solutions like this, but to a certain extent they all miss the bigger picture. The problem of akrasia is the problem of how to characterise *rational* practical reasoning in general. Before we can claim that actions are *irrational* we have to be able to show in what sense action is typically rational and even to do this we have to figure out what kind of authority reasons for action possess. These are the real issues that have to be addressed, and so starting by assuming some kind of internalist view and then figuring out how to consistently avoid or solve problems of irrational action is never going to be a really satisfactory strategy.

In keeping with our attempt to uncover the internalist assumptions found underlying many solutions to akrasia it might be useful to take a brief detour into the origins of motivational internalist views.

### 1.2 Action and motivation: an outline of the internalist perspective

As we have seen there is a distinction between those who think motivation is internal to obligation (or evaluation more generally), and those who think it is external. *Internalists*, as they have come to be called, argue that there is a necessary connection between judgements about what to do and the motivation to act on those judgements. Having an all-things-considered judgement about what to do is considered by internalists to be sufficient to motivate an agent to action. So, if we are wondering whether, for example, Mary will come to help us out at the charity fund-raising event, all we need to know is that she really thinks that, all-things-considered, helping out is what she should do. Given that this is genuinely the case we know that she will be motivated to act.[20] This is not to be confused with keeping one's word. What we know about Mary is something stronger than that she always keeps her promises. Rather we know that once she judges that all-things-considered, she ought to act in a certain way she *will* act.

But there is a sense in which this seems to be almost counter-intuitive. In ordinary life we often find ourselves *wanting* to do something irrespective of whether we think it is best; indeed, it is not the usual case to find ourselves being motivated to do something purely because we judge or believe that is what we *should* do. Phenomenologically at least, it appears as if there often seems to be a push from some feeling or other that is independent of our judging that such-and-such is the best thing to do. This could suggest that we require some sort of special and completely independent interest in doing what we believe we should do.

Our common intuitions about action thus do appear to support the claims of the *externalists* who hold the view that the connection between judgement and motivation

[20] It is important to note here that we are assuming that Mary is *able* to come and has not changed her mind. So long as there are no last-minute crises to attend to, she hasn't broken her leg, lost her car keys etc. and we know that it is her all-things-considered judgement that she should come, then we can assume with confidence that she will come.

is not a necessary one, and rather that the motivation or interest to act must be present along with the judgement or reason that this is the right act to perform. If that is the case then how is it that internalism seems to be a commonly held position on action-production? It is really not possible to answer this question or to see how it is that the problem of akrasia has so often been dissolved rather than resolved without a very brief review of the history of how obligation came to be conceived as internal to the agent's own moral or generally evaluative reasoning.

### 1.3 The rise of internalism: a brief sketch[21]

Normativity, in particular *bindingness*, or the requirements or demands of morality that are binding on all rational agents as such became a particular issue for seventeenth-century ethics. The notion of morality as the recognition of binding demands developed into a view of morality that was radically different from the ancient view of morality as being fundamentally about the good. For the ancients, all action is aimed at some good and it is the task of ethics to find out what the final good is. General deliberation about what to do was thought to make sense only in the context of the agent's good, so for an action or activity to be something an agent ought to do is the same thing as furthering the agent's good. For Plato, but not for Aristotle, the good is person-neutral – good reasons for action are ultimately determined by *the* Good and rational action aims at the Good. The good, according to Aristotle, is the good for each particular person themselves, and so deliberation only makes sense in relation to the

[21] Stephen Darwall, *The British Moralists and the Internal Ought 1640-1740*, pp.1-22 (New York: Cambridge University Press, 1995); JB Schneewind, *The Invention of* Autonomy (London: Cambridge, 1998). I follow Darwall's and Schneewind's discussions of the shift from the ancient to the modern concept of morality closely here.

good of the agent – the eudaimonia or flourishing life of the agent.[22] Rational action then, was necessarily connected to the good of the agent and actions were only considered binding if they furthered the good of the agent. Given that this was the case there just was no concept of morality as binding *requirements* or *demands* on all rational agents as such. In effect, there was no motivational internalism in the sense that we have become familiar with in modern times.

The beginnings of modernity in the seventeenth-century marked the start of the rejection of Aristotle's teleological metaphysics. It could no longer be universally assumed that an action an agent ought to do is the same thing as an action that furthers the agent's good. But if this is rejected then there is a problem – different people might have different goods. Without the kind of teleological metaphysics of Aristotle that *metaphysically* guarantees a harmony of goods across individuals it becomes possible that the good of one person could conflict with the good of another.[23] So what should an agent do if aware of this? However, this question only makes sense if what one ought to do is different from seeking one's good. If obligation is stripped of its foundation in teleological metaphysics in this manner then the task of the moral philosopher changes radically. The task of the moral philosopher is not to find the final good, but first and foremost to figure out what normative status itself *is* if it is not necessarily linked to the good, and thus uncover the rational foundations of those requirements that bind us.

The modern conception of morality began its development from the seventeenth-century natural law tradition of Grotius, Pufendorf, Cumberland, Hobbes and Locke. The initial impetus motivating these thinkers was their reaction against the

---

[22] Aristotle, *Nichomachean Ethics*, I.1-2, trans. W.D. Ross (Oxford: Oxford University Press, 1980).
[23] Contemporary virtue ethicists however, argue that we do not need a teleological or perfectionist view of human nature to establish an Aristotelian style virtue ethics. Virtue ethics can be conceived of as a normative view about which characteristics or traits are worthy of moral esteem in a moral agent. We do not have to look to an ideal of human nature in order to have a view of what is worthy of such esteem. See Stephen Darwall (ed.) (2003).

tradition started by the Stoics and developed formally by Aquinas. According to Aquinas, natural law (or morality) follows eternal law and eternal law is the paradigm of divine wisdom which moves "all things to their due end" (that is, it determines the ideal state of every natural thing).[24] Natural law is a formulation of eternal law made accessible to rational beings and specifies norms or laws which apply to all rational beings as such. The good of each being is its perfection. So what I ought to do is to realise my own good (and this involves not preventing others from realising theirs – as eternal law would have it).[25]

Grotius in *The Law of War and Peace* (1625) presents a very different picture:

> There is no law of nature [in the Thomistic sense], because all creatures...are impelled by nature towards ends advantageous to themselves...[C]onsequently, there is no justice, or such there be, it is supreme folly, since one does violence to his own interests if he consults the advantage of others.[26]

But this objection to Thomistic tradition can only work if it is possible for people's interests really to conflict and if what is required by natural law does not always and necessarily further a particular agent's own interests. Grotius thinks that people's interests *can* and *do* differ and that the natural law sometimes requires us to make sacrifices for the good of society or others. But we would not even be able to formulate this view on a Thomistic or Aristotelian picture.[27] Grotius presents us with a world in which one has a basic desire for one's own good and for society's good in so far as it is advantageous to oneself, and unlike on the Aristotelian picture, these *can* conflict. Natural law is there to help us figure out what to do when our interests are conflicting with those of others.

---

[24] *Summa Theologica,* XCIII, 1, in Anton C Pegis ed., *Basic Writings of Saint Thomas Aquinas*, (New York: Random House, 1945). Quoted in Darwall, p.5.
[25] *Summa Theologica,* XCIII,1-3. Darwall, p.6
[26] *The Law of War and Peace.*, Prolegomena, pp.10-11, trans. Francis W. Kelsey (New York: Carnegie Endowment for International Peace, 1925). Quoted in Darwall p.6.
[27] It is interesting to note that Plato was acutely aware of the possibility of such conflict. This is evident especially from Book 1 of the *Republic*.

But now the obvious question arises – why should we go on assuming that there are norms which are binding for all rational persons that require them to sometimes act against their own interests? If we deny, as Grotius does, that an agent's own good is the source of her obligations and that the goods of different individuals cannot ultimately conflict, then what justification do we have for claiming that being an action an agent ought to do is the same thing as being an action that furthers her good? The question now facing Grotius is why we should assume that any rational person is obligated to act against their own good. If the natural law is not grounded in the good of the agent then from where does it get its status as obligating or rationally binding? Grotius responds to this scepticism by moving towards the idea of morality as *self-governance*, claiming that even if God did not exist we would still be under binding obligations to do whatever makes social life possible.[28] To be an action that an agent ought to do is not identical with that which furthers her good, but neither is it identical with obedience to divine commands. Rational agents direct their own actions without interference from the state, God, peers, threat of punishment etc. Hence rational agents must be capable of *moving themselves* to act on the recognition of their moral obligations. So here we can see the beginnings of what we know as internalism; in so far as we are rational beings we are self-governing agents who are motivated to, and will act on what we believe to be our obligations even if we believe we are required to act against our own interests in favour of the interests of society in general.[29]

From Grotius' rejection of the Thomistic tradition onwards there was a gradual shift away from understanding obligation as linked to the good of the agent towards the idea that obligation can only be properly understood if we know how it is linked to self-

---

[28] *The Rights of War and Peace*, p.xix, trans. anon, from Barbeyrac's French translation (London, 1738). For an in-depth analysis of the role of Grotius in the development of the modern concept of normativity see J.B. Schneewind, *The Invention of Autonomy*, chs. 4-5 (Cambridge: Cambridge University Press, 1998).

[29] Grotius almost moved in a direction more akin to enlightenment views on motivation and action but his concern with the authoritative role of the moral law holds him back (unlike Hobbes).

governance or autonomy. This is not to say, however, that there was uniform agreement on this. In fact both Hume and Hutcheson denied that self-governing moral agency is necessarily connected to obligation. For Hume, for example, it is not the agent's capacity for self-governance that is necessarily connected to obligation. But agents do direct their own actions because it is their own *motives* to act that are in some manner necessarily connected to obligation. What we can see from the differences among these early modern moral theorists is that one can be an internalist by claiming either that obligation is necessarily linked to the capacity for autonomy or self-governance (reasons internalism), or, that obligation is about the agent moving herself to act and she can only do this if she has the relevant *motives* to do the action in question (motivational internalism). The idea that a moral agent moves herself to act as she judges best or right is what unites these theorists as internalists. But with the rise of empirical naturalism the same basic internalist concept developed quite differently within two diverse traditions.

The empirical naturalists (Hobbes, Cumberland, Hutcheson, Hume, Locke) wanted to account for obligation in natural terms.[30] Nothing "can superinduce a Necessity of doing or forbearing any thing, upon a Human Mind deliberating upon a thing future" other than motives, so obligation consists in motives that are unavoidable for agents who deliberate rationally.[31] Empirical naturalists were anxious not to posit powers of reason that go beyond what is necessary for empirical enquiry.[32] So all the empirical naturalists held that there are no *sui generis* or special powers of practical reason, but rather the latter consists in just those motives for action which arise either directly from

---

[30] Although Locke did have God as the supreme authority sanctioning morality, he nonetheless still explained the normative force of morality as internal to the agent. *An Essay Concerning Human Understanding.* Hobbes *Elements of Law* (1640), *De Cive* (1642); Richard Cumberland, Bishop of Peterborough, *A Treatise of the Laws of the Laws of Nature,* (1672) trans. John Maxwell (London, 1727).
[31] *A Treatise of the Laws of Nature, p.*233 (see Darwall p.15).
[32] But Cumberland in particular did not have a simple instrumental view of reason. He claimed that there is also theoretical reasoning which can lead us to new ends (e.g. universal benevolence).

natural sentiments or from sentiments that follow upon the exercise of ordinary, factual, theoretical, and reflective uses of reason. Those such as Locke, Hutcheson and Cumberland argued that a calm reflective use of theoretical reason could lead agents with psychology such as ours to new ends even though reason could dictate no ends by itself. This is different from the instrumental reason internalism of Hobbes who argued that obligation consists in doing what is necessary to achieve the goals that determine our deliberation (and for Hobbes this was self-preservation).[33]

The second major internalist tradition centres around the claim that morality concerns the self-determining or self-governing moral agent (the first centred on motivational internalism). Obligation consists in the motives to act gained through the exercise of autonomous practical reasoning. There are two main strands of this view: One strand of the autonomy view of obligation is focused on the idea of *accountability.* The concern here was to try to establish what the connection is between obligation and accountability. An agent can only be held accountable if she is a being who can bring herself to act as obligated. Locke tries to show how a moral agent is only obligated to do what she herself would choose to do through exercising her autonomous practical reasoning. So all obligation really is is a conclusive motive for acting that arises through autonomous practical reasoning. Self-determination or self-governance involves stepping back from present desire and using theoretical reason to vividly consider the objects of what one desires and in this way one comes to a judgement about which course of action would provide desire satisfaction. Hence autonomous practical reason really is just the same as the use of theoretical reason in that like theoretical reason, practical reason cannot motivate by itself, being merely instrumental it just allows the agent to consider clearly the objects of her desires. Locke gets his

[33]Another different type of theory about this time were the moral sense theories of Hutcheson and Hume. For these obligation is about the kind of response one has to traits or motives perceived in other agents. This is a slightly different sort of internalism.

influence here from Cudworth for whom ethics (i.e. obligation) must be grounded in *this* sense in pure practical reason. Being moral is about following a "law of love" where pure intellect is motivated by universal love. If we are able to draw on this when deliberating about what to do then we will be motivated to act according to our deliberations. This same theme runs through Locke and Cudworth who both stress the capacity to stand back from current desires and draw upon standing motives we might have and to make a motivating judgement of our own good on the whole. It is only because we are imperfectly rational that we need to exercise this capacity of standing back. Autonomous agents determine their conduct by careful reflection and by making judgements about their own good. This is all they need to guide conduct. Crucially, they do not need to believe that desire for their own good *obliges* them to act.

The second strand of the autonomy view of obligation concerns not accountability but *authority*. On this view the autonomous agent is self-governing. The agent must be able to distinguish between the 'power' and the 'authority' of her motives.[34] What we see emerging here with Butler and also Shaftesbury (and later Kant) is the development of an understanding of the conditions necessary for the very possibility of an autonomous will. Reason must be purely practical, and what this means is that for an agent to have an autonomous will she must be able to form a normative judgement that is intrinsically motivating. On the accountability view above, for Locke and Cudworth the agent does not have to recognise that the motive she has for action is somehow *authoritative*, that is, as one on which she should act. On their view if the agent was to fully exercise her power of autonomous practical reasoning then she would simply find this motive compelling (i.e. most likely to further her good or happiness) and this is what makes it obligating.

However, for Shaftesbury and Butler on the authority view, it is a necessary

[34] Joseph Butler, *Fifteen Sermons Preached at the Rolls Chapel* (London, 1726) Sermon II, para.14.

condition of autonomy that the agent herself be guided by some sense of her motive being authoritative were she to fully exercise her capacity for autonomous practical reason. What this means is that in order for an agent to be self-governing it is necessary that she step back from her current desires and consider both the objects of her desires and her independently existing motives to act and make a judgement about what action will further her happiness: it is this stepping back that is the exercise of autonomous practical reason. For Shaftesbury and Butler it is also necessary that she recognise that this judgement is conclusive, that is, one on which she should act. It is not enough that the judgement itself is intrinsically motivating. This then is what being obligated to act is about. This normative theory of the will is developed later into the concept of the will as the ability to make demands of *oneself.*

### 1.4 Internalism and the explanation of weak-willed action

From even the above very brief survey we can see that internalism has a history that accounts for its status as more than just a stipulative claim that judgement and motivation are necessarily connected. It has become traditional and even trivially true to claim that, for any rational agent, to judge that an action ought to be performed is intrinsically motivating. It takes no great step to see the prima facie difficulties caused by akrasia. If an agent's *ceteris paribus* judgement is that she ought to Q then, she will Q. But of course the weak-willed agent fails to Q. So strong has been the internalist conception that akrasia has continually been subject to (often radical) redescription in order to render it consistent with the necessary connection between judgement and motivation. As we have discussed above problems of practical irrationality generally

have been resolved by this sort of redescribing of the phenomena. There are two main types of redescription:

1) the claim that akrasia is not solely about acting against better judgement but is more precisely described as a situation where the agent does not have or form the right sort or *appropriate* intention to act (Davidson, Hare, Bratman);
2) the claim that overwhelming desires of some sort put the agent in a situation of lack of self-control to varying degrees. Thus akrasia is not strictly about *intentionally* acting against better judgement because the agent does not intend to perform the action as it is akratically described. According to this perspective weak-willed action may either be intentional but not voluntary or not intentional at all. The most important feature here is that weak-willed agents lack some particular ability or disposition that is as important in the production of rational action, if not more so, as the ability for practical reasoning (Mele, Jackson, Stocker, Pugmire, Watson, Pears and Austin).

What these resolutions have in common from the standpoint of both classical and ordinary or intuitive descriptions of free will is that they seek to accurately analyse weak-willed action so as to make the possibility of such action consistent with a version of internalism.

There are a couple of important points to note here. As we mentioned above it is almost universally accepted (that is, by internalists as opposed to externalists) that problems of practical irrationality such as akrasia can and need to be made consistent with internalism. If it is true that having, holding or making a judgement is intrinsically motivating then we need an explanation for those cases which appear to falsify this altogether. The very term '*practical irrationality*' belies an answer to this problem. The agent who fails to act as she judges best all-things-considered is somehow not properly rational. This explanation can be cached out in many different ways but the

basic idea remains the same: rather than seeing these anomalous cases as counterexamples to internalism itself we can persist with our internalist theory of judgement and action if we put weak-willed action into the category of irrational actions. Weak-willed agents do something wrong in their practical reasoning about what to do and there are numerous theories as to what this is, but they all can be subsumed under the two broad resolutions I have outlined above. Weak-willed agents are irrational because if they were fully rational they would be motivated to act as they judge best all-things-considered and, *ceteris paribus,* they would not act contrary to their best judgement all-things-considered. While it is undeniable that weak-willed agents are in some sense irrational the problem with leaving the matter at that (so to speak) is that we are in danger of simply assuming that if we were rational we just would act as we think we should. However as we will see later, this assumption is difficult to support and it is weak-willed cases that bring this to the fore.

## 1.5 Strong Internalism[35]

Both strong and weak motivational internalism maintain the view that holding a moral judgement is necessarily or intrinsically motivating. So an agent holding a moral judgement must have at least some degree of motivation to act in accordance with it. The degree of motivation depends on how strong one's internalism is. For strong internalists, the agent has sufficient motivation to perform the act in question, but weaker versions of internalism deny this. We can also say that the agent has some

[35] Stephen Darwall (1995) calls this version of internalism "existence internalism" and this in turn is what David Brink is referring to when he discusses what he calls "agent internalism" (Brink, 1989). Some confusion is unavoidable here and since there does not appear to be uniformity across the board regarding the use of these terms I have cut the cloth to my own shape with what I am calling 'strong' and 'weak' internalism.

*reason* to act. It is pertinent here to distinguish between two types of reasons: motivating reasons and normative reasons. This is a crucial distinction to recognise since it is quite clear that some reasons for action are motives but not all motives are reasons for regarding an action as obligatory, right or justified. For example, Jane's cat has five kittens but the landlady is adamant that her tenants can have only one pet each. Not wanting to break the conditions of her lease Jane takes the kittens and abandons them in a box outside a shop two streets away. Jane's action is intelligible. She has a motive for her action and she acts upon it – we can call this a *motivating* reason. But no one with any concern for animal welfare would argue that what Jane did was justified or right since she could easily have secured the well-being of the kittens by making a few calls to animal welfare organisations. So we can say that Jane had no normative reason for what she did because at least according to the common standards of her community[36] abandoning helpless animals to their own fate is considered cruel and wrong behaviour. If we suppose that no motivation is entailed by holding a moral judgement we can also imagine similar scenarios where an agent could have a normative reason without having any motivating reason to act. We can at least say that it seems as intelligible to hold that agents can have normative reasons to act without motivational reasons as it does to hold that agents can have the latter without the former.

Strong motivational internalism, however, is the denial of this distinction. It is the assertion that by virtue of holding a normative reason to act the agent also has a motivational reason for acting. The argument here is that real normative reasons must be practical and they can only be practical when they can actually motivate the agent to act accordingly. It is not possible then for an agent to hold a normative reason for

[36] We do not have to argue that these normative standards are in any way objectively true moral facts about the world. The distinction between normative and motivational reasons does not depend on any type of moral realism. Normative reasons need only reflect agreed or commonly accepted standards.

action without also holding a motivational reason. Thomas Nagel articulates this position very clearly. He states:

> The first-person acknowledgment of a sufficient reason for doing something...is sufficient to explain one's doing it...There is no need to hold that 'X acknowledges a reason to A' *entails* 'X does A, or X wants to do A'....All I wish to claim is that such an acknowledgment is by itself *capable* of providing a motivation in the appropriate direction – that there is no need to seek an alternative or supplementary explanation of action when that one is available.[37]

The strong internalist position is fairly clear.[38] The motivation is internal to any self-addressed moral judgement or, in general, normative reason that the agent considers obligating. However the motivation the agent has is not merely a consequence of holding a normative judgement, rather it is characterised as a conceptual truth that the motivation is intrinsic to the holding of the moral judgement. The motivation is sufficient for the action in question; but if the agent fails to act as she thinks she should then she is weak-willed or practically irrational, and weak-willed agents have difficulties either forming the right sort of moral judgement in the first place or with their powers of self-control. By allowing for agents to be moral, except when they are not, strong internalists increase the plausibility of their position. R.M. Hare is one notable exception here who argues that akrasia as commonly conceived could not exist given the sufficient motivation provided by one's moral judgement.

A powerful intuition giving force to the strong internalist position is the idea that an agent can have a normative reason to do something only if she can intentionally do it; and if she cannot be motivated to do it, then she cannot do it. So if holding a

---

[37] *The Possibility of Altruism,* (Oxford: Clarendon Press, 1970), pp.110-111.

[38] I think Nagel causes some confusion in this passage by not stating what he seems obviously committed to. To clarify, saying merely that it is a conceptual truth that acknowledged reasons by themselves are *capable* of motivating an agent raises the question when are they sufficient and when are they not. That is, Nagel seems to hold back from saying that reasons are necessarily motivating. But that seems to render it mysterious as to when a reason "by itself" motivates and when it does not. It seems to me that he must be committed to the stronger claim (i.e. the conceptually necessary connection between the acknowledged reason and the motivation) even though in this passage he seems to placate common sense by denying such an entailment. The discussion of Nagel here interprets him as expressing this stronger claim.

normative judgement[39] entails having a moral or normative reason to act accordingly, then the motivation to act must be able to be present.[40] This of course, is a conceptual claim. A real normative reason is practical, or concerned with what to do. But if a moral judgement entails a normative reason then this means that the judgement must be motivating since it can only be practical when it actually has the capacity to motivate. So having a normative reason to do something implies motivation to do it because the only way to have a normative reason is to hold a moral judgement and holding a moral judgement is motivating.[41] On this view then, if an agent judges that she should *A* in circumstance X, then she is motivated to *A* in X.[42] This conceptual connection is sometimes expressed as a requirement of rationality such that if it is considered right or justified (either objectively or by commonly accepted standards of rationality) for an agent to *A* in X then there is a prima facie reason for that agent to *A*. In other words, to judge that A is right means that it is a reason for the agent to act, and reasons are rationally binding imperatives which entail motivation, as we explained above. We can also say then, on this view, that an agent has a reason to act in a certain way if she would be motivated to act in that way if she were practically rational. So, an agent who judges that she has a reason to act but is not motivated is *practically irrational* – there is something amiss with her process of forming an intention to act based upon her

---

[39] It is usually understood here that the moral judgement must be at least justified if not true. If we allow that any moral judgement no matter how bizarre made by any agent is a normative reason to act then this view becomes implausibly strong. Once again we have to argue that to be plausible this view must maintain that only justified moral judgements are normative reasons to act and it is these judgements that entail a capacity to act accordingly and be motivated to do so.

[40] See Robert Audi, *Moral Knowledge and Ethical Character* (Oxford: Oxford University Press, 1997), pp.232-233.

[41] Sometimes this is expressed by saying that having a moral reason implies having *some degree* of motivation to act. However I believe this is a somewhat imprecise characterisation. If we take it that each moral judgement provides us with a *prima facie* reason for action, as this view does, then it would be more accurate to say that having a moral reason provides us with *sufficient* motivation to act if, when we act, we act upon our all-things-considered or best judgement.

[42] John McDowell, (1978) 'Are Moral Requirements Hypothetical Imperatives?', *Proceedings of the Aristotelian Society* Supplementary Vol. pp.13-29. (1979) 'Virtue and Reason', *The Monist*, pp.311-350. (1985) 'Values and Secondary Qualities', in Honderich (1985). Thomas Nagel (1970), *The Possibility of Altruism* (Princeton University Press). Mark Platts (1979) *Ways of Meaning* (Routledge and Kegan Paul). (1981) 'Moral Reality and the End of Desire' in Mark Platts ed., *Reference, Truth and Reality* (Routledge and Keegan Paul).

judgement of what is best to do.

The crucial point that these strong internalists are making is that what we really mean when we say it is right for an agent to *A* is that this is a fact about the agent's reasons for action. So, to judge that action *A* is right is to say that there is a reason to do *A* and reasons are simply requirements of rationality. For all rational agents, then, to recognise that one has a reason to act is to be motivated to do so. Now this motivation will be *prima facie* motivation and so will not necessarily be acted upon. If Jane recognises that she should give to charity then she has a reason to do so. But if her two children are hungry then she may decide that her obligation to give to charity is overridden by her obligation to feed her family and so she will be motivated to act accordingly. In this sense then, simply having a *prima facie* reason for action does not provide sufficient motivation to act. It is one's *all-things-considered* judgement about what to do that provides the sufficient motivation to act.

It is important to note here that my distinction between weak and strong internalism does not pick out, as the principal distinguishing feature, the *degree* of motivation that is internal to making or holding a moral judgement. But it is increasingly common for this feature to be focussed on as the locus of the distinction. David Brink, for example, determines the difference as follows. Weak internalism about motives is the claim that it is a conceptual truth that moral considerations provide *some* motivation; but strong internalism about motives is the claim that it is a conceptual truth that moral considerations provide *sufficient* motivation for action.[43] However, as I have argued, what distinguishes weak from strong internalism is not the degree of motivation entailed by holding a moral judgement but rather it is the extent to

[43] David O. Brink 'Externalist Moral Realism', *Moral Realism and the Foundations of Ethics* (Cambridge, 1989), pp.39-50. Robert Audi *Moral Knowledge and Ethical Character* (Oxford, 1997), pp.217-237. David Copp 'Moral Obligation and Moral Motivation', in Jocelyne Couture and Kai Nielsen eds., *On the Relevance of Metaethics: Canadian Journal of Philosophy* Supplementary Volume 21 (1995), pp.187-221.

which various forms of practical irrationality – self-deception, akrasia, over-whelming desire etc. – are considered to be possible and consistent with the connection between judgement and motivation. Weak internalists, as we will see below, do not characterise the connection between judgement and motivation in such strong terms as to make practical irrationality impossible or inconsistent with internalism.

Let us take it as a fact about motivation that a change in motivation follows a change in normative judgement. Even externalists will not deny that there is a reliable connection between change in motivation and change in judgement. What they deny is that this connection is a necessary one. Now internalists seeking to explain the admitted reliability of this connection must explain another fact – that agents act according to their normative judgement except when they do not, as it were; in other words, that most of us act, for example, morally only some or most, but not all of the time. But what about the occasions when we fail to act morally (or as we value best)? Let us assume it is true that agents can only act morally if they recognise a moral obligation as self-addressed.[44] Given this the internalist not only needs to explain the reliability of the connection between judgement and motivation but also the occasions on which agents recognise that, all-things-considered, they have a moral obligation to act a certain way but nevertheless fail to act accordingly. Whether an internalist is a strong or weak internalist is best determined by the extent to which they can allow for this possibility. As we will see however, it is weak internalists only who describe occasions on which we fail to act as we think we should as simply occasions *on which we fail to act as we think we should.* For strong internalists the conceptual connection between moral judgement and motivation is not one which leaves room for akrasia or

[44] The agent recognises that she has an obligation to act which is binding. It is important to note here that we are not discussing situations where the agent knows that such-and-such is *expected* of her by society or her peers, but where she herself does not recognise the obligation as binding. For example, John knows that it is expected of him to volunteer to work in the soup kitchen but he does not agree that this is a binding obligation *on him*. This does not make John a hypocrite, he just does not accept that this is what he ought to do.

self-deception as traditionally conceived. Nevertheless the fact remains that ordinary rational agents are moral or act according to their moral judgements except on those occasions when they appear to do quite the opposite. It has been suggested in the past that the existence of weak-willed action can provide a successful argument *against* strong internalism. We know that weak-willed action occurs so surely this must falsify the strong conceptual connection favoured by the internalists?

Hare pre-empted this objection in his defence of prescriptivism in the 1960's: if moral judgements are prescriptive then it would be impossible to accept or hold a moral judgement and act *contrary* to it. Considering this sort of objection, Hare quotes Hume where he says "'tis one thing to know virtue, and another to conform the will to it"[45]; people are always doing what they think they ought not and therefore on this basis we could argue that prescriptivism is wrong. Hare's response however is rather simple: if moral judgements were not prescriptive there would be no problem about moral weakness, but there is a problem about moral weakness so therefore moral judgements are prescriptive.[46] The key to Hare's solution to this difficulty is to focus on the nature of what he calls 'moral weakness', and this has been one of the most popular strategies to adopt when trying to make akrasia consistent with strong internalism. The most plausible avenue for a strong internalist to take here is to conceive the phenomena as a case of overwhelming desire or some such thing and this re-describes the practical reasoning of the weak-willed agent in such a way as to avoid a head-on collision with thestrong internalist conceptual connection between moral judgement and the motivation to act, a connection the most important feature of which is its *indefeasibility*. However as we will see redescriptions of this kind tend to reconfigure the phenomena out of existence.

---

[45] *Treatise,* iii.I.i, (Hare p.67).
[46] R.M. Hare 'Backsliding', from *Freedom and Reason* (OUP, 1963), p.68.

As we mentioned briefly above there are two main types of redescription of akrasia and strong internalists tend to favour the first option: (1) the agent is said not to have or form the right sort of judgement that would necessarily be connected to her motivation (for example, Davidson, Bratman), or the agent is said to believe herself to be acting against her better judgement when in fact she never had that better judgement in the first place and so akrasia involves a degree of self-deception (Hare); (2) overwhelming desires put the agent in a situation of lack of self-control and so akrasia is not strictly about *intentionally* acting against better judgement (Mele, Jackson, Stocker, Pugmire, Watson, Pears and Austin). Weak internalists prefer this second option. However, before we examine what I am calling the attempted redescriptions (their proponents would not, of course, see them as such) in greater detail, let us get clear about what weak internalists are claiming.

### 1.6 Weak Internalism

Like strong internalism, weak internalism takes the connection between holding a moral judgement and being motivated to act to be a reliable connection, but unlike the stronger position, this connection, while still considered to be a necessary one is not thought to be *indefeasible* in light of cases such as akrasia or self-deception. There is one particular version of weak internalism that has become popular in recent literature, which I shall term 'rational agent internalism'.[47] The claim is as follows: if an agent judges that it is right to do A in circumstance X then either she is motivated to do this

[47] The other type is consequential internalism – the kind of weak internalism that I think is all that can be plausibly maintained.

or she is practically irrational.[48] This means that if an agent judges that she ought to X then she will necessarily be motivated to X unless she is disrupted by akrasia or self-deception. So, in so far as the agent is *rational* she will be motivated to act and *will* act as she judges best (all-things-considered). It seems implicit in the very idea of a rational agent that such an agent would not make moral judgements without taking them seriously at least to the extent of tending to act on them when nothing else interferes with their judgements. But the strength of this position comes from a view of the rational agent as having *desires* that are in line with their moral beliefs about what to do rather than from any view of what is logically necessary about the internality of motivation to moral judgement.

It is worth noting again here that weak internalists are not best distinguished from strong internalists by the degree of motivation they consider to be internal to holding a judgement; so the point is not that strong internalists claim that the motivation provided by holding a moral judgement is *sufficient* to cause the agent to act accordingly whereas weak internalists claim that only some or *prima facie* motivation is internal. Rather the difference between these two positions is most clearly determined by the extent to which each position accounts for a weak-willed agent's actions. Weak internalists do not redescribe akrasia to be some kind of failure of reasoning, analogous with a failure of theoretical reason, as strong internalists do. But they *do* explain akrasia as a *practical* failure of rationality or as a failure on the agent's part to be a fully rational agent at the time of the action. These two reactions to weak-willed action do *not* amount to the same thing. Weak internalism is consistent with akrasia defined as action against one's better judgement, whereas strong internalism finds that definition to be conceptually impossible. But since it cannot be denied that

[48] Michael Smith calls this the 'practicality requirement' on moral judgement. Michael Smith *The Moral Problem* (Blackwell, 1994), p.61. Internalists of this type include Smith (ibid.); Mark Johnston 'Dispositional Theories of Value', *PAS* Supplementary Vol., (1989), pp.139-74; Philip Pettit and Michael Smith 'Practical Unreason', *Mind*, (1993), pp.53-79.

agents act against their own self-confessed better judgement strong internalists believe that akrasia needs to be redescribed as some sort of failure to come to one's proper better judgement in the first place.

The weak internalist then is quite as committed to the necessary connection between holding a moral judgement and having the motivation to act as the strong internalist. If we take the ideal of practical rationality to be the intention to act according to the results of our deliberations then when this *fails* to happen strong internalists argue that the breakdowns (akrasia, self-deception etc.) that occur must be categorised as failures of rationality. These are typically portrayed as involving some sort of deficiency, that is, some sort of ignorance, error, inattention or illogic – all resulting in a failure to form an intention properly. The important point about these failures is that they are essentially cognitive failures somewhat like breakdowns of theoretical reason. Weak internalists share the same ideal of practical reason as strong internalists but characterise occasions when we fail to intend according to the results of our deliberations as failures of *practical* reason, such as a failure of autonomy or of self-control.[49] But what theoretical commitments have weak internalists made that entails such a view of akrasia?

It is not any kind of belief about the degree of motivation entailed by holding a normative judgement that leads to this view of akrasia as a practical failure. Rather what does lead to this position is a certain view of the agent. Strong internalists, for their part, take on the burden of arguing for a conceptual connection between judgement and motivation that is true for all agents irrespective of their individual

[49] Frank Jackson 'Weakness of Will', *Mind* (1984), Alfred Mele, *Autonomous Agents* (OUP, 1995); Gary Watson 'Skepticism about Weakness of Will', *Philosophical Review* (1977); Michael Stocker, 'Akrasia: The Unity of the Good, Commensurability, and Comparability' from *Plural and Conflicting Values* (OXford: Clarendon, 1990); David Pugmire 'Motivated Irrationality Part I', *Proceedings of the Aristotelian Society,* 1982.

dispositional characters. Agents will form an intention to act based on their best judgement and this best judgement will be, one might say, *akrasia-proof* – once the agent forms their best or all-things-considered judgement there will be action in accordance with it. But the only solution to the problem of akrasia on this view is to redescribe so-called weak-willed agents as somehow not forming their best judgements properly, and this is the source of their 'weak-willed' actions. So this, in effect, amounts to a denial of the possibility of strict weak-willed action – akrasia is not really action against one's best judgement all-things-considered but is action that results from the failure to form an intention to act that *is* in accordance with one's strongest reason (best judgement all-things-considered) for action.

Weak internalists on the other hand do not take on the same burden – they do not try to argue that the connection between judgement and motivation to act is necessarily true for all agents. Instead they argue that this connection is necessarily true for all *rational* agents. Now there is a notable distinction here. One might indeed ask what the difference is between the strong internalist claim that rational agents will act on their all-things-considered judgements but it is recognised that sometimes they do not do so, and the weak internalist claim that there is a connection between holding a moral judgement and being motivated accordingly that holds for all rational agents only? Are these not essentially the same – both claiming that the existence of akrasia is not inconsistent with internalism? The difference is this. The strong internalist position amounts to the claim that we must assume that all agents will, ceteris paribus, act according to their best judgement about what to do and that if they do not do so then they are somehow lacking in the necessary ability to form the right intention based on their best judgement. The weak-willed agent's most crucial error concerns their practical *reasoning*. However weak internalists from the outset assert that their claim only covers agents in so far as they are rational. This is important because it means that

they are not forced to postulate that the failure of the weak-willed agent is *necessarily* a quasi-logical or inferential one, as strong internalists are. Once it is claimed by the strong internalists that all agents come to act by deliberating and forming an intention to act based on the results of their deliberation then when the appropriate or expected action fails to occur we are committed to maintaining that the agent's error is in some sense a logical one – they somehow have failed to form the intention that corresponds to their best judgement all-things-considered. If the agent's error is a logical or inferential one then all other descriptions of their error, such as irresistible desire, seem not to pinpoint the real problem with practical reasoning the agent is having. The most we can do is try to give some explanation for why the agent could have come to make such an error. But such explanations are not satisfactory by themselves since they give no indication as to the nature of the crucial error – they only serve to suggest possible reasons why an agent occasionally falls off the logical wagon, so to speak.

Weak internalists however are not precluded from offering wider explanations of weak-willed action. The claim that rational agents are constructed in such a way that they necessarily tend to act on their best judgement all-things-considered does not commit weak internalists to maintaining that the only relevant error weak-willed agents make is a logical one concerning the formation of correct and appropriate intentions to act. For there is some plausibility in saying that moral agents are, more often than not, rational agents and as such their desires are in line with their considered judgements about what to do. But this is only true on the assumption that the degree of rationality required to hold a judgement in the first place implies that there must be a connection between judgement and motivation which is sufficient to produce the motivation to act according to the judgement held. When an otherwise rational agent fails to act as she judges best the weak internalist can subsequently offer a broader range of explanations; for since the facts about the rational agent are contingently established we are not

forced to defend the logical possibility of some version of akrasia as strong internalists are. Strong internalists can only make a weaker version of akrasia consistent with their internalism, but weak internalists can account for the possibility of strict weak-willed action.

What we need to do at this point is to take a closer look at just how these different types of internalism resolve (or dissolve) the problem of weakness of will. The next chapter is given over to this task.

# Chapter Two

## Weakness of will and Motivational Internalism (2): the Solutions

### 2.1 The strong internalist solutions to the problem of weakness of will

Strong internalism has been an influential position on the connection between holding a moral judgement and being motivated to act since Plato, and it is this view, I believe, that sustained interest in weakness of will until the arrival of weak internalism on the philosophical scene appeared to offer a readily acceptable compromise to the problem. Socrates in the *Protagoras* rejected the possibility of weakness of will outright since he held the view that human beings are psychologically so constituted that they freely do A at t only if they do not, at t, take some competing action to be better. Plato's later position in the *Republic,* is not significantly different from the Socratic one except that he admits the possibility of action against better judgement but denies that this is *freely* done. The principal question for Plato is whether human psychology actually *permits* strict akratic action. We will not deal with Plato's views at any length save only to point out that his view that strict akratic action[1] is inconsistent with the necessary and sufficient connection between holding a judgement and being motivated to act on it for every properly ordered soul, has persisted through to contemporary strong internalists. Their problem is, as we have seen, a very similar one – how to make weakness of will and internalism, two seemingly inconsistent things, consistent with one another. Just to recall, if weakness of will really is action against better judgement and internalism claims that holding such a judgement is sufficient motivation for action then we must either redefine or reject one of these concepts because they cannot both be true together.

[1] That is, akratic action that is defined as action against one's better judgement all-things-considered.

As we mentioned above there are two common ways to analyse weakness of will:

1) the claim that weakness of will is not really about acting against better judgement but is a situation where the agent does not have or form the right sort of judgement that would necessarily be connected to her motivation (Davidson, Hare, Bratman); or the claim that weakness of will involves some degree of self-deception where the agent believes herself to be acting against her better judgement when in fact she never had that better judgement in the first place (Hare).
2) the claim that overwhelming desires of some sort put the agent in a situation of lack of self-control and so weakness of will is not strictly about *intentionally* acting against better judgement (Jackson, Stocker, Pugmire, Watson, Pears and Austin).

What both attempts have in common is that they are used to show that weakness of will is not really action against one's better judgement about what to do. This denial of the possibility of strict akratic action is the most common strategy strong internalists employ to eliminate any inconsistencies they may stumble upon. For the sake of simplicity and clarity I think it would be best not to try to canvass all of the specific arguments of those strong internalists with an interest in weakness of will, but rather to focus on this cross-section of strategies corresponding to these two most common ways such theorists attempt to analyse weakness of will.

## 2.1 The 'Usurping Reasons' View

Weakness of will occurs when an agent acts upon reasons that are not her strongest reasons for action – we will call this the 'usurping reasons' view. The most well-known proponent of this strong internalist (type 1 above) view is Donald Davidson. He argues that the akratic agent intentionally acts against her better judgement but she still acts for a reason even though her reasons are not her strongest or all-things-considered reasons. Davidson's description of akratic reasoning goes as follows. The agent judges that *all things considered* it would be better *not* to Q and believes that she will not Q. However she also has another judgement that it would be better *to* Q and this judgement is not relative to any other particular considerations the agent has, it is just an unconditional judgement or evaluative attitude. What distinguishes that akratic agent from the continent agent is that the former ends up acting on this judgement despite having a better reason to avoid doing so.

An important commitment that Davidson makes is to deny any distinction between *wanting* most and judging or *valuing* best. He maintains that *all* desires are expressed by value-judgements[2] and so judging a course of action to be best (all-things-considered) means that I will pursue it and that I *want* to do it. But the converse also holds. So, if an agent performs an action we can assume that (1) she wanted or desired to do it, (2) she judged it better to perform that particular action than any other available alternative, and (3) that the action performed was actually accompanied by this (unconditional) judgement that the intended action is better than any alternatives. Consequently the akratic agent is not some strange species of agent who is not tied to these constitutive norms of intentional action. However what is immediately obvious here is that if the agent is acting against her better judgement (as indeed she must be if

[2] 'Intending', (1980) p.98.

she is to be called akratic at all) then the problem or the error in practical reasoning which gave rise to the weakness of will in the first place must be located elsewhere than between her judgement or conclusion that the course of action she performs is best and her actually performing the action. One was of putting this is that weakness of will for Davidson, concerns incontinence of belief or judgement rather than incontinence of action. An akratic agent's *action* is always performed intentionally because it is always performed with the belief or judgment that the action is the best (or best one available). So Davidson suggests that the real problem here is not the agent's action but her belief or intention *formation*. Since her intention to act is formed from an unconditional judgement that stands alone, as it were, disconnected from other relevant considerations that might outweigh it, the agent ends up acting intentionally yet contrary to her own best judgement. This is what weakness of will must be according to Davidson, and it is the only possible way to describe what happens when an agent acts against her better judgement and also remain consistent with strong internalism as we have characterised this view.

Michael Bratman[3] is another strong internalist who, like Davidson, wants to defend the usurping reasons view of akratic action; but he thinks that for this view to be plausible we have to change the Davidsonian conception of the conclusion of a practical inference. Bratman argues that Davidson cannot drive the necessary wedge between the agent's best judgement and her practical conclusion, so he suggests that the form of the conclusion is really something like 'I shall do *a*' because this actually settles the question of what to do. Davidson, on the other hand, only gets as far as '*a* is desirable'. Hence for instance if an agent, Alf, drinks wine akratically then he does so

[3] Bratman, Michael (1979), 'Practical Reasoning and Weakness of the Will' *Noûs*. Bratman's views appear to have altered somewhat in recent years. His view on weakness of will does not seem to be significantly different, however he now argues that a proper analysis of weak-willed actions shows that there are problems for those who argue that practical reasons fully determine intentions to act. See Bratman (2003).

against his better judgement and so he must have two flatly contradictory conclusions – 'I shall do *a'* and 'I shall not do *a*'. He finds drinking desirable and not desirable for the same reason (that it induces drunkenness). But a problem remains with this analysis – how does the course of action get settled if the agent is stuck between equally convincing judgements? Bratman thinks that this will not arise because there is a norm of rationality that governs these transitions from the evaluative proposition to the conclusion about what one will do. However, despite their different views on the content of a best judgement and the nature of intentions to act both Davidson and Bratman think that the necessary connection between holding an all-things-considered judgement and the corresponding motivation to act is self-evident in the sense that it is not possible to reach a practical conclusion and *not be motivated to act on it*. They both maintain, in essence, that akratics reach practical conclusions also, but their conclusions do not have the content they would have had, had they been fully rational.[4]

There are problems with the usurping reasons view that Davidson, and to a greater extent Bratman, struggle to answer. A common criticism is that acceptance of the connection between judgement and motivation is based on a confusion or failure to make a relevant distinction. For example, Elizabeth Rapaport argues that Davidson "fails to distinguish between wants in the sense of particular desires and wants in the sense of rational preferences or over-all self-interest."[5] David Pears (1982, 1984), Gary Watson (1977) and Richman (1980) put forward similar arguments. Another common strategy used here is the attempt to identify a true principle which this necessary connection between judgement and motivation resembles in an attempt to explain why a seemingly false principle such as the strong internalist one might seem true to some philosophers. These objections raise significant difficulties for the usurping reasons

[4] That is, their practical conclusions about what to do are not based upon their best or all-things-considered reasons for action.
[5] Rapaport (1973) p.180-1.

view of weakness of will; but what I want to suggest is that there is another difficulty with this view which may indeed preclude the possibility of a usurping reasons view of akratic action from the outset.

This difficulty is caused by certain of the key assumptions that the strong internalist requires in order to maintain her position. These assumptions are:

1) All-things-considered judgements about what to do are action-guiding.
2) Agents always act intentionally.
3) To form an intention to act is to form an unconditional judgement about what it is all-things-considered best to do.
4) If an agent judges that X is better than Y and the agent is free to do either X or Y then the agent wants or is more motivated to do X rather than Y.
5) If an agent wants to do X more than Y and is free to do either then she will do X if she does either.

While not every strong internalist accepts all of these assumptions, this is, broadly, the range of assumptions to be found. What I want to suggest here is that holding any of assumptions 3-5 leaves no other option but to describe weakness of will as a case of inferential failure.[6] This description may cover some cases of what we agree to be akratic action but it does not cover all cases. So, if strong internalism cannot satisfactorily account for all types of akratic action then there must be a problem with some of the basic assumptions, some of which may have to be rejected as a result.

Now assumptions (1) and (2) I will just leave aside since neither claims anything that any internalist – strong or weak – would reject. So let us take it as

[6] There is a possible ambiguity concerning inferential failure that is worth noting here. This could be interpreted as (1) failure to make a certain inference (that is, one which ought to have been made), or as (2) making a logically unwarranted inference (that is, an invalid one). (1) amounts to a failure to infer whereas (2) is inferring incorrectly. However in cases of practical reasoning I think we can see that (1) and (2) are the same. The agent who fails to infer the correct conclusion about how to act from her reasons for acting is making the same error as the agent who infers the wrong conclusion from her reasons.

uncontroversial that having a judgement that X is the best thing to do is action-guiding and it is similarly uncontroversial that when agents act they act intentionally since to act intentionally just is to freely will oneself to act. But what about assumption (3)? This assumption is one which is not accepted by all strong internalists. Davidson's approach for example, assumes that an all-things-considered judgement will always be of the form 'all-things-considered, action *a* would be best', but Michael Bratman argues to the contrary that an agent may perform an action based on an all-things-considered judgement and yet not be prepared to say that the chosen course of action was *best*. Davidson argues that the akratic agent thinks that it is *best* to, for example, have a third glass of wine when she knows that all things considered, she should stop after the second.

This is quite puzzling for two reasons. Firstly, it could be considered unrealistic. Agents do act with the judgement that it is, say, "all right" to have the third drink despite their better judgement but they do not always think that it is *best* to have a third drink. We can think of many similar cases to this. Secondly, it does appear to be unnecessary for the strong internalist to continue to insist on this. As David Pears points out, it is enough that the action be accompanied by a *permissive* evaluative judgement. To be a strong internalist, then, one does not need to argue that the agent contemplating drinking the third glass of wine judges that the third glass is best because the reasoning can still work in exactly the same way with the claim that the agent thinks the third drink is *all right* or *acceptable* – i.e. permissible.[7] Our strong internalist can still argue that the problem lies *before* the formation of the practical conclusion about what to do and not between it and the action. However the reason why Davidson believes that he must argue that the agent has the judgement that her proposed action is

[7] See Pears (1984), pp.187-189.

best is because he considers and rejects the assumption that there is a distinction between wanting to do something because it is *best* and wanting to do it *without* judging it best to do it. It seems that Davidson here is committing himself to the view that reason is instrumental in the sense it is not possible to want to do something without thinking it to be best where 'best' really just means 'most wanted'. If an agent's goal or desire is to drink more wine then reason takes that goal or desire to be the thing or state of affairs most wanted and figures out how best to attain it. This is, of course, a Humean account of action and Davidson adds to this by arguing that all desires are expressed by evaluative judgements.[8]

But if the Davidsonian position is right then we are left with the consequence that if an agent intentionally A-ed (that is, judged A unconditionally best) then she must have intended to A. In other words, judging best is a *necessary condition* of acting intentionally. But a problem arises for this position if we can successfully argue that akratic action against an intention to act (i.e. an unconditional best judgement about what to do) need not involve the agent's having competing intentions. Judging an action best,[9] it could be argued, is neither a necessary nor a sufficient condition for the performance of that action. If the process of forming an intention to act just is the process of forming an unconditional best judgement about what to do then akratic action can only be explained as a type of inferential failure. But if we agree that there are cases of akratic action that do not seem to be about a clash of competing intentions (best judgements) then there must be something wrong with assumption (3) to begin with because we will be able to show that not all intentions to act are unconditional judgements that X is *best*.

The following is an example of an akratic action from Alfred Mele[10] that does

---

[8] 'Intending', p.86.
[9] Along with the physical and psychological abilities to perform the action.
[10] Mele (1987), p.19.

not appear to be about competing intentions to act. Mary judges that it would be best to shoot an injured horse to put it out of its misery. While aiming at its head she notices the sad expression in its eyes and, *still thinking that it would be best to shoot it*, she intentionally refrains from doing so. But she has not reversed her judgement about what would be best to do. So, if she freely refrains from shooting the horse and has not 'changed her mind' (i.e. her judgement) then this is an akratic action. But Mary does not require competing intentions, according to Mele. When she intentionally refrains from shooting the horse she has not changed her *judgement* about what it is best to do but she no longer has the *intention* to shoot. If this is the correct description of at least some types of akratic action then there is an implication. If the all-things-considered judgement that it would be best to shoot can remain even when the agent no longer *intends* to shoot, then we must conclude that all-things-considered judgements are *not* intentions. What we can say, then, and what examples like these show, is that there are cases of weakness of will when it is true to say that the agent fails to intend according to her all-things-considered judgement.[11]

Consider another example. John wants to draw blood with a needle for a biology assignment. He judges all-things-considered that this is best and forms the correct intention (i.e. the unconditional judgement that he should prick his finger). But when he brings the needle to his finger he fails to pierce the skin. On Davidson's analysis John could not be akratic because he intends according to his all-things-considered judgement. It would seem, however, that the best explanation of this case is that John is akratic; but it also seems that agents who manage to intend according to their all-things-considered judgements are not akratic at all. So how can we explain

[11] It is interesting to note that Davidson clarifies his position on this in his 'Replies' in Vermazen (1985) p.206. He makes it clear that weakness of will is really about the failure to form conative propositional attitudes in the right way. Intentions are one type of such attitude. So weakness of will is not about an error in belief, but it is about the failure to form these attitudes in a "rational, coherent way". By this I suppose he means failure to form an intention in accordance with one's all-things-considered judgement.

why John fails to act on his unconditional judgement about what it is best to do?

Mele argues that Davidson's understanding of continent action could be shown to be at best inadequate and at worst false in light of cases of akratic action such as John's above (i.e. action against *unconditional judgement*). Davidson argues that if it were true that akratic action directly contradicts continent action then his causal theory of action would be in trouble, since if it is possible for an akratic to act against her unconditional judgement (i.e. her intention) about what is best to do then akratic action is in direct opposition to continent action – action in accordance with one's unconditional judgement, which is in turn formed in accordance with one's all-things-considered judgement about what to do.[12] This would in effect mean that there is no guarantee that agents will act in accordance with their intentions to act and so strong internalism would be in jeopardy. To see off this threat the strong internalist has to argue that it is possible to act akratically even though I have already formed the intention to act otherwise and that actions like this do not, in fact, threaten internalism. To prove this however it is crucial to show that Davidson is fundamentally wrong about the process of forming an intention to act. However what *is* threatened by the possibility of action against intention is strong internalism of the type we have been discussing.

If it can be shown that the process of forming an intention to act cannot be what Davidson thinks it is then it will be clear that assumption (3) of the internalist's basic assumptions listed above is false. What might also be clear at this stage is that it is only strong internalists who would entertain such an assumption. We will be dealing with this more directly elsewhere but what we need to see at this stage is that forming an intention cannot be as simple as Davidson thinks it is.

---

[12] Hence his concern about always acting with the judgement that what we are doing is best and his pressing the point that the 'flaw' in the akratic's 'reasoning' takes place elsewhere than between the practical conclusion and the action etc.

An attempt to do this is made by Bratman who argues that it is because Davidson fails to appreciate the significant co-ordinating role of future-directed intentions that he fails to see the extent to which his theory is in danger of being impaled upon the horns of two classic problems: agglomerativity and Buridan's Ass. Buridan's original problem was a consideration of how an ass, between two piles of hay judged equally desirable, could intentionally go to one over the other. Bratman's version of this problem asks whether an agent, who knows she can stop at only one of two equally desirable bookshops on the way home from work has intentions to go to each shop or neither. She judges unconditionally that any act of stopping at Kepler's would be just as desirable as any act of stopping at Printer's Inc. What intention does Davidson's agent have?[13]

According to Bratman unconditional 'desirability' judgements are implicitly comparative. However there are two types of comparisons. A *weak* comparison would see going to Kepler's as *at least* as desirable as any alternatives, whereas a strong comparison would see it as *strictly* more desirable. Davidson, Bratman believes, would presumably say that his theory requires the latter. If a weak comparison is all that is required then the agent will intend to go to Kepler's and also intend to go to Printer's Inc. despite knowing that she cannot go to both. This goes against a natural constraint on intentions – that they be agglomerative – which, Bratman assumes, Davidson would not care to violate. If an agent rationally intends to go to X and rationally intends to go to Y then it should be possible for the agent to rationally intend to go to X *and* Y. However if we claim that weak comparisons are all that are required for future intention then intentions will not be agglomerative. So, clearly Davidson would not argue for *weak* comparison or evaluation because of the threat to agglomerativity.

---

[13] Michael Bratman, 'Davidson's Theory of Intention', *Two Faces of Intention* (Cambridge: Cambridge University Press, 1999), p.219.

Yet if he chooses to argue for *strong* comparisons, as Bratman presumes he must, then Davidson will be faced with a rather intractable Buridan problem. If Davidson argues that our bookshop agent only has an intention if she has a strong evaluation or comparison in favour of one option then she has neither intention because, like Buridan's ass, she finds both equally desirable. But, as Bratman points out, the agent will simply *decide* which bookshop to visit while continuing to see each as equally desirable. She might just decide to go to Kepler's even though she does not have an unconditional evaluative judgement that this course of action is strictly more desirable than going to Printer's Inc. So the intention the agent acts on is not arrived at in the manner specified by Davidson's theory of intending.[14]

Without even examining assumptions (4) and (5) we can see that any strong internalist is in significant difficulty concerning the nature of intentional action. Examining these would take us further away from the tasks at hand. The principal difficulty these assumptions cause for understanding and explaining weakness of will is that they commit us to a denial of what is often called 'clear-eyed weakness of will'. Davidson's view, for example, does not allow for the possibility of what has come to be called 'clear-eyed' weakness of will. An akratic agent acts with 'clear-eyed' weakness of will if she acts with the judgement that what she is doing is against her better judgement. Since an akratic agent for Davidson supposedly always acts with the judgement that that particular course of action is best, 'clear-eyed' weakness of will is not even a possibility. So wherever the problem or mistake is located in the agent's practical reasoning it is never found between the practical conclusion and the action. Take the case of an agent who buys chocolate despite having judged it best to buy ice-cream. When he buys the chocolate the agent must believe, if Davidson's view is

[14] ibid., pp.220.

correct, that what he is doing is best despite the fact that he also has a judgement that all things considered, buying ice-cream would be best. But it just does not seem correct to say that an agent in this situation believes that buying the chocolate is *best*. Davidson attracts quite a bit of criticism for this position on the type of weak-willed action possible and indeed this position is a consequence of assumption (4) in particular, hence the focus on the self-evidence Davidson finds regarding this principle. The main difficulty this account of intentional action encounters is in defending the tight connection between motivation and evaluation in (4). Denying this connection opens up a whole array of possible explanations of weak-willed action, from failure of self-control due to motivational force of irresistible desire, to lack of alignment between evaluations and desires.[15]

Given the description of acting on an intention strong internalists prefer, they consequently only really have one option available to describe the reasoning an akratic agent goes through before acting, and that is the 'usurping reasons' view suggested by Davidson. This view of akratic action is best described as a form of inferential failure on the part of the agent. Instead of forming her intention or best judgement on the basis of her all-things-considered judgement she fails to take that last final step. Of the cause of this failure, whatever it is, we can at least say that it is not a justifying reason for her action. So insofar as she fails to take the final step necessary to the performance of the intentional action the agent is stuck in a process of what is best described as inferential failure.[16]

To sum up the argument so far, if we agree that there are cases of akratic action that do not quite fit this picture (basically any case where the agent appears to act against and *with* an already formed intention; also cases where the agent can act

[15] Much of the dispute with Davidson centres around offering these sorts of alternatives to (4). See in particular Alfred Mele (1987); David Pears (1984); Robert Audi (1979); Gary Watson (1985); David Pugmire (1982).
[16] See footnote 6.

without thinking X the *best* thing to do) then it follows that we have to admit that these cases cannot be explained without arguing that the agent does not really have the right intention in the first place. But this is an unsatisfactory and problematic solution because these are significant counter-examples that cannot be dismissed. This shows then that there is a serious problem with the strong internalist assumptions. We saw that (3) is problematic enough to cause a serious problem for internalism (type 1). Assumptions (4) and (5) are problematic too.[17] So now we can see that Davidson's strong internalism not only fails to provide a plausible solution to the problem of weakness of will, but that it is a problematic theory of intentional action. It has been the examination of puzzling cases of apparently intentional but akratic actions that bring these problems to the fore. If it is possible to have an intention and then act against it then the connection between having a judgement and acting could not be as tight as has been hitherto argued.

## 22 The 'No Genuine Judgement' View

This position is in many ways closely related to the Platonic position articulated in the *Republic* because the distinctive feature of this view is that the weak-willed agent is characterised not as acting on a reason outweighed by her better judgement but as failing to assent to or recognise the obligatoriness of the judgements they make. In the contemporary debate this view is associated with R.M. Hare. His view that the weak-willed agent does not have a genuine judgement to begin with seems to amount to nothing more than the view that weak-willed agents like those in our examples are

[17] See footnote 62.

either insincere about what judgements they *do* in fact hold or that they are simply deceiving themselves. So we should really say that such an agent would like herself to judge prescriptively that she ought to X (where X is the course of action approved of as her best judgement); but thinking that 'it would be nice if I was really committed to doing X' does not seem to amount to a commitment to doing X. It is difficult to see how this type of solution (or, the 'no genuine judgement' view as I am calling it here) can satisfy us as an explanation of weakness of will. But since this is a familiar strategy it is worth taking a look at how Hare attempts to defend this position.

Since it is not possible, on Hare's view, to make an evaluative or normative judgement and *not* recognise that this is what one ought to do, akratic agents cannot be said to make proper evaluative judgements in the first place.[18] Akratic agents do not just act against their all-things-considered best judgement in favour of a lesser judgement, rather they never had an all-things-considered judgement, and they have no *genuine* judgement about how best to act.

Hare argues that it is part of the meaning of moral terms that judgements containing moral terms are intended as guides to conduct (that is, are intrinsically motivating). But how is this to be reconciled with the obvious existence of akratic actions? Surely if moral judgements were prescriptive then it would be impossible to hold a moral judgement and act *contrary* to it? Hare argues that not all moral judgements have the same prescriptive force that the "perfect specimen" has.[19] These "off-colour" moral judgements are different because they do not imply 'can'. So it is possible to say 'I ought to but I cannot'. This is not absurd, rather the agent is just using 'ought' in an off-colour way. Weakness of will is a special case of this. Hare argues that there are two ways to solve the problem of moral conflict (of which

[18] R.M. Hare (1963).
[19] Ibid., p.68.

weakness of will is an example for Hare): (1) to recognise that wanting or desiring is not the same as making a value-judgement; or (2) to accept that moral judgements are universalisable. But Hare recognises that he cannot simply define 'value-judgement' in such a way that if an agent did not do what she thought she ought she could not be using the word 'ought' evaluatively. This would be stipulative and artificial, leaving no possible way to explain akratic acts other than to say that the agent does not actually have a proper understanding of what she means when she says that she has a moral judgement. Hare seems faced with two dead-ends – deny weakness of will altogether or allow its existence to falsify his prescriptivism. Instead Hare chose to allow for the possibility of weakness of will by building some defeasibility conditions into his view. The akratic agent still can be said to have no genuine judgement, but weakness of will is just that which prevents the agent from forming her all-things-considered judgement properly.

Let us look closer at Hare's problem, which leads him to account for weakness of will in this way. According to Hare there is a close connection between saying 'I ought' and 'wanting'. When we want something what is going on is very different from what goes on when we think some descriptive proposition is true. Wanting or seeking is more like those kinds of thoughts classified by prescriptivism as choices, resolves or requests. Value judgements on the other hand are *universalisable* and this is what makes them differ from desires. Wanting is like assenting to a singular imperative[20] and not to a moral or other value-judgement. Making up my mind what I want to do is very different from deciding what I *ought* to do because ought involves considering whether anyone else placed in similar circumstances would be expected to act as I act. Thinking about what I *want* to do only involves consulting my own

[20] ibid., p.71.

desires. It would be much easier to decide what we ought to do if moral judgements were *singular prescriptives* of some sort (i.e. singular values such as desiring).

But Hare argues that if we accept instead that such judgements are universalisable and reject the universal prescriptiveness of moral *principles*[21] then it is easier to decide what to do because when I say I ought to do A what I mean by that is that I and anyone in similar circumstances ought to do A. There are lots of things we would say that we ought to do if we did not think we were committing ourselves to a prescription (and therefore to an action). The real difficulty is finding an action one is prepared to commit oneself to and which is also universalisable. The action-guiding roles of principles and judgements are different on this view. An agent can accept a principle without being prepared to act on it whereas the agent's own moral judgement about what she ought to do is a *commitment* to do the action in question. But this judgement only has moral force if it is universalisable. Being weak-willed, or 'backsliding' as Hare calls it, means making a moral judgement without committing oneself to the prescriptivity or universalisability of it. Moral language has lots of ways to avoid the pure prescriptivity or universalisability built into it and this is just what is going on when agents act akratically. The true moral question involves asking myself what action I am prepared to commit myself to and at the same time accept as exemplifying a principle of action binding on anyone in like circumstances.[22] It is perfectly possible for an agent to commit themselves in this way and yet fail to act (e.g. getting out of bed). But it does not mean that they do not really accept the principle, instead their moral strength just fails them. They still mean the same thing but just do not have the strength to carry through.

Another way of backsliding is special pleading.[23] Here the agent starts off

---

[21] ibid., p.73.
[22] ibid., p.74.
[23] ibid., p.76.

accepting a principle as binding; but when she considers that it might be against her own interests she stops prescribing to herself but continues to think that it should apply to everyone else (but this universality is only *descriptive*) or at least those whose interests do not concern us. The word 'ought' still retains all the *descriptive* meaning it had but stops expressing a universal prescription. These are just two ways the expression 'I ought' can lose its robustness without departing from its original meaning. This amounts to an inability to realise our ideals, according to Hare, and this has been termed weakness of will, moral weakness, etc.

However, Hare still believes that it is evident that we cannot sincerely assent to a command addressed to ourselves and *at the same time* not intend to or be motivated to perform it, if now is the occasion for performing it and it is in our physical and psychological power to do so. So typical cases of moral weakness are situations where the agent says that she 'ought but cannot' and the 'ought' here is less robust because it does not carry the same prescriptive force – the prescription for the *particular* case has been withdrawn but the agent may still say that it applies to actions in similar circumstances.

To summarise Hare's view: a moral judgement entails a prescription which in turn embodies a desire, and to have this desire is to be disposed to act accordingly. To fail to act is only possible if one's assent is merely verbal (insincere) or if one fails to understand the language one is using. But the weak-willed agent does not act according to her judgement and therefore on Hare's theory the weak-willed agent cannot have the corresponding desire and so cannot accept the imperative and does not accept the moral judgement. The judgement is not genuine.

Hare's attitude to weakness of will is a consequence of his rather complex

views about meaning, that is, about what it *means to say* that an agent holds a moral judgement. Without getting involved in a critique of these views, however, we can still see that even if Hare is right about the prescriptivity of moral judgement, this solution to the problem of weakness of will fails to move us beyond the puzzle we began with. Let us take Hare at face value and accept his view of what it means to hold a moral judgement. Now let us take an example. Mary has the judgement 'I ought to feed the cat'. Holding this judgement entails the prescription (the bindingness of the judgement) which in turn embodies a desire (to feed the cat), and to have this desire is to be disposed to act accordingly; so Mary will go and feed the cat provided something beyond her control does not prevent her along the way. 'I ought to feed the cat' becomes an expression of the agent's decision about how she should act. But if Mary fails to act then we are back at the same old worry that troubles strong internalists – it seems irrational, impossible even, that practical reasoning leading to a conclusion about feeding the cat should result in not feeding the cat. Telling us that Mary never really or genuinely had the judgement that she should feed the cat in the first place rings hollow as an explanation of what is going on. The notion that the akratic agent does not have a genuine judgement to begin with seems to amount to nothing more than the view that agents like Mary are either insincere about what judgements they *do* in fact hold or that they are simply deceiving themselves and so we should really say that Mary would like herself to judge prescriptively that she ought to feed the cat. But as we have seen already, thinking that 'it would be nice if I was really committed to doing X' does not amount to a commitment to doing X. It appears then, that given these difficulties, the problem of weakness of will comes no closer to being solved on the 'no genuine judgement' view.

## 2.3 The weak internalist solution to the problem of weakness of will

So far then we have seen that both the strong internalist attempts (that is, the 'usurping reasons' view and the 'no genuine judgement view') to solve the problem of weakness of will are problematic to the extent that one may justifiably wonder whether there is any solution to be found by continuing in that direction at all. Weak internalist solutions do at least show an awareness of some of the problems we have encountered with strong internalist solutions so far. I will confine this section to discussing the weak internalist solution to weakness of will as a generic position. However I do not want to give the impression that there is unanimity here – there are significant differences to be found among those who I characterise as weak internalists. In what follows then I shall treat the views of Alfred Mele as representative of what I am calling the weak internalist solution to the problem of weakness of will but it is important to keep in mind that this is more a strategy of illustrative convenience than an accurate presentation of a uniformly held position.[24]

Mele considers his own position to be a weaker internalism than Davidson's and argues[25] that agents *can* act akratically against best judgement. Mele[26] argues against Davidson in particular, that there is no species of judgement that is immune from weakness of will and that having a best judgement is neither necessary nor sufficient for intentional action. Most crucially, he believes that agents require a capacity to exercise *self-control* in favour of performing the action they consider to be best all-things-considered.

Central here and common to most weak internalists is a firm rejection of the view that Davidson's 'unconditional judgements' are intentions. On this alternative

---

[24] Alfred Mele (1986, 1987, 1995). Other weak internalists include David Pugmire (1982), Elizabeth Rapaport (1973), David Pears (1982, 1984), Robert Audi (1987), Simon Blackburn (2000), Frank Jackson (1984), William Frankena (1976), Robert Dunn (1987), Michael Stocker (1990).
[25] Mele (1995).
[26] Ibid.

picture the ideally self-controlled agent will *come to want* what she decides is better all-things-considered. But a less self-controlled agent may not be motivated to act according to her all-things-considered judgement *even if* she has that judgement and the corresponding intention to act.

Weak internalists argue that the logical problem concerning the possibility of *intentionally* acting against better judgement is not the real problem of weakness of will – the real problem lies elsewhere. Akratic actions may be free in the sense that there is no strong compulsive desire present pushing the agent to act, but this does not mean that the akratic agent *intends* to act akratically. The agent who knows, for example, that by walking in the snow she will get her shoes wet and walks in the snow does not *intend* to get her shoes wet – she just intends to walk in the snow. Actions intended under one description may not be intended under another.

If we examine the strong internalist's argument in terms of *intentions* to act rather than unconditional best judgements in favour of acting we begin to see the extent of the claim they would have us accept. If an agent intentionally A-ed (i.e. judged A unconditionally best) then she must have intended to A. In other words, judging best is a *necessary condition* of acting intentionally. Mele however, wants to argue that akratic action against what he calls a 'here-and-now' intention need not involve the agent's having competing intentions (*even if* doing X intentionally entails intending to do X).[27] Judging an action best (along with the physical and psychological abilities to perform the action) is neither a necessary nor a sufficient condition for the performance of that action. The crucial move the weak internalist makes against the strong internalist is to deny that forming an all-things-considered judgement followed by an unconditional judgement in favour of the proposed best course of action is the process

[27] Mele (1995) p.19.

of coming to have an intention to act.

Mele argues that there are cases of weakness of will when it is true to say that the agent fails to intend according to her all-things-considered judgement, but he asks whether it is possible for an agent to act akratically although they form the correct intention to act (that is, they intend according to their all-things-considered judgement). The above example of walking in the snow is supposed to convince us that it just might be possible to act akratically even though one has formed the correct intention. Consider again the example of John and the needle we met above.[28] Strong internalists have difficulty with this case because John seems to still intend to pierce the skin (that is, he intends in accordance with his all-things-considered judgement) and this would make him unakratic or continent. However John still seems to be akratic. Strong internalism seems to have no resources for dealing with cases where the agent forms the correct intention but nevertheless fails to act.[29]

Mele argues that the strong internalist's understanding of continent action could be shown to be at best inadequate and at worst false by akratic action that is like John's. According to Davidson for example, continent agents act on their *un*conditional judgement that X is best but akratics do not get as far as forming the right intention or unconditional judgement. However if there really is a type of akratic action that is against intention (i.e. unconditional best judgement) and if Davidson is right about intention being an unconditional judgement, then this means that akratic action appears to pose a threat to causal theories of action.[30] Davidson does seem to think that if akratic action directly contradicts ordinary or continent action then his theory of action

---

[28] John wants to draw blood with a needle for a biology assignment. He judges all-things-considered that this is best, forms the correct intention (i.e. the unconditional judgement). But when he brings the needle to his finger he fails to pierce the skin.

[29] We will be discussing this type of case in more detail in Chapter Five.

[30] How can an agent act on a reason that is not her strongest reason?

is in trouble.[31] To see off this threat Mele argues that it is possible to act akratically even though I have already formed the intention to act otherwise and that actions like this do not, in fact, threaten causal theories of action. To argue this point, however, Mele needs to show that Davidson is fundamentally wrong about the process of forming an intention to act.

In Mele's view, the problem lies in supposing that the balance of the agent's motivations must always be in line with her here-and-now intentions. So John's action in the example above seems inexplicable because we assume that he is most motivated to do what he intends to do. The balance of motivation may shift as he attempts to execute his intention. His desire not to harm himself may be supplanted by a weaker desire to prick his finger and we can give a causal account of this as a reason for action. It is obvious, says Mele, that John's motivation may be out of line with his evaluation of the object of his desire. An agent can rank her wants in the *evaluational* sense (a ranking of their *value* – what we evaluate are the objects of our wants, and these have motivational force) but this does not have to imply that she is most motivated to do what she ranks highest.[32] Mele wants to argue that the "motivational force of wants typically is not under our control to the extent to which the evaluative ranking of wants is."[33]

Just because evaluative judgements are action-guiding does not mean that the judgements themselves are logically or causally sufficient for the presence of the corresponding intentions. Judgement can be broken down into two elements: the content of the judgement and the judging attitude itself. The question that must be asked here is whether the attitude of judging-best entails the motivation to act

---

[31] Hence his concern about always acting with the judgement that what we are doing is best and his pressing the point that the 'flaw' in the akratic's 'reasoning' takes place elsewhere than between the practical conclusion and the action etc.
[32] CCW Taylor makes a similar suggestion in Taylor (1980).
[33] Mele (1995) p.39.

according to the content of the judgement? Robert Dunn[34] argues that evaluative judgement is theoretical and non-volitional and that an evaluative best-judgement can be dissociated from an agent's motivation. Similarly Mele wants to distinguish between a best *belief* that A-ing is best and a best *judgement* (best belief is the genus, best judgement is a species). Best *judgemenst* rationally commit agents to action on pain of practical irrationality.

So, John ranked his competing wants and formed his intention to act accordingly, but the motivational force of his competing want at the time was not in alignment with his ranking (maybe because of a scary childhood experience involving needles). He was less motivated to prick his finger at the time he formed his intention to prick it. However Mele still ends up wanting to know why the agent acts on her lesser reason rather than her stronger or all-things-considered reason for action.[35] Things become slightly more difficult for Mele here because he also accepts (with Davidson) that an adequate reasons explanation of an action must explain *why* an agent did X *instead of Y*. This is pertinent to weakness of will because we want to know why the agent acted as she did instead of as she judged best or intended to act. Mele goes on to argue that proponents of motivational internalism should *not* be committed to maintaining that the reasons an agent takes to be most important are the cause of the agent's action. In some cases (such as John's) the action is not caused by the reason he takes to be most important. However Mele wants to reject altogether the Davidsonian view of unconditional judgements. It is not the case that John judges unconditionally that refraining from pricking his finger is best. He does not take himself to have made

---

[34] Dunn (1987) p.124.

[35] p.39. There is a suggestion here from Mele that an explanation of this is not available to the agent herself– the agent knows *that* she acted on her lesser reason but not *why* she did so. (p.40) However I am not convinced that weak-willed agents are suffering from self-deception to such an extent. To say that an explanation is not available suggests that even if they thought about it they would not be able to understand why they acted as they did, and this, surely, does not seem to be a true feature of weakness of will.

this judgement. The ideal agent will, unlike John, always have the balance of her motivations in accordance with her intentions. But to suggest that anything less than this is weakness of will is too extreme.

The akratic agent must in some way have the balance of her motivations out of line with her judgement of what is best to do. Mele accepts this but rejects the view that the cause of this imbalance must be some form of unruly desire. Take any case of supposedly strict akratic action. Is it true to say that the agent *freely* acts intentionally against her better judgement? Why not simply argue that she is overcome by irresistible desire and hence is acting against her will? Mele is anxious to argue against a strand of argument that sees desire as irresistible if all the agent's resources of resistance fail.[36] Mele sees a 'crucial flaw' in this type of argument, namely, the premise that there is nothing more the agent could have done by way of resistance. Mele argues that there are plenty of things the agent could do in a situation where there is some strong contrary motivation, for example, refuse to focus attention on the pleasantness of the akratic act, concentrate on the benefits of acting according to best judgement, promise herself a reward for acting continently, etc. These are 'self-control techniques' and Mele points out that there are even more sophisticated ones such as behaviour modification therapy. To ignore all these possible "non-deliberative modes of resistance" is to be unaware of the fact that "measures such as these are plainly capable of having a salutary influence on the balance of one's motivations at the time of action."[37]

So, according to Mele, what we have traditionally understood as the problem of weakness of will is really a matter of sorting out some prima facie paradoxes of self-control. Continent action is the result of an agent's self-control. If an agent is self-

---

[36] Similar arguments are put forward by Wright Neely, 'Freedom and Desire', *Philosophical Review* 1974, and David Pugmire, 'Motivated Irrationality', *PAS*, 1982.
[37] Mele, (1995) p.24.

controlled then she will evaluate her wants in light of other wants and relevant beliefs and she will be motivated to act accordingly. That is, she will come to want what she decides is better all-things-considered. A less self-controlled agent may not be motivated to act according to her all-things-considered judgement *even if she has that judgement and the corresponding intention to act* (p.45).

It is crucial for most weak internalists that unconditional judgements *not* be identical to intentions. This link has to be broken in order for self-control to be relevant to the explanation of weakness of will. If making an unconditional judgement (in accordance with one's all-things-considered judgement) was considered to be the formation of an intention to act then we have to explain why it is that the agent failed to act on the intention she had formed. Here we *do* get into issues of rationality or consistency. But if the agent has unconditional judgement *without* the intention to act then it is easier to appeal to other factors to explain why she is not acting according to her better judgement. It would not make any sense to appeal to a lack of self-control if the agent failed to act as she *intended* because to fail to act as one intends is to fail to act according to how one's has decided to act and we would need more than 'lack of self-control' to explain this.

### 2.4 What weak internalism really commits us to

Let us begin here by asking what it is that weak internalists seek to achieve. The quick answer to this, of course, is that they want a connection between reason and motivation that is not quite as strong as strong internalists prefer to characterise it. What I intend to show in this section, however, is that whereas weak internalism offers more promise,

it ultimately is still too strong. However it *is* possible to supply a connection between reason and motivation that does the job internalists want it to do without going so far as to claim that there is some feature of normative judgements that motivates the sincere holder of them.

Most internalists characterise the modality of the connection between judgement and motivation as that of necessity, since most internalists argue that internalism articulates a sort of conceptual truth (i.e. motivation is conceptually intrinsic to holding a moral judgement). As we have seen weak internalists also argue that the connection is one of necessity but not to the extent that makes internalism indefeasible and therefore implausible. However it is only when we restrict the connection between judgement and motivation to fully rational agents that internalism seems entirely plausible, for it is implicit in the concept of a rational agent that they do not make moral judgements without taking them seriously at least to the extent of tending to act on them when nothing else is at stake. We could say that rational agents are so constructed that they necessarily tend to act in accordance with their moral judgements.

There have been some interesting examinations of the internality of motivation to judgement in recent years, and one recurring point is that what is often taken to be the result of a strong internal connection might be better understood as the result of how rational agents *happen* to act when they take what they consider to be the moral point of view. Robert Audi argues, for example, that the plausibility of internalism does not derive from the internality of motivation to normative judgement but derives instead from the plausibility of the view that rational agents must have desires that are in line with their normative beliefs about what they should do. Internalism of this kind,

he argues, is strong enough to maintain the view that it is necessary for anyone who holds a moral judgement to have some degree of motivation to act accordingly. This is because *rational* agents always have some degree of motivation to act and not because the connection between judgement and motivation is conceptually intrinsic.

In short then, weak internalists argue that the logical problem concerning the possibility of *intentionally* acting against better judgement is not the real problem of weakness of will. Weak-willed actions may be free in the sense that there is no strong compulsive desire present pushing the agent to act, but this does not mean that the weak-willed agent *intends* to act in a weak-willed way.[38] Actions intended under one description may not be intended under another.

But as we shall see, weak internalism argues for a connection between judgement and motivation that is still too strong. Here is an example from Audi:[39] a deontologically oriented father judges that it is his moral duty to punish his son for taking the car out without permission by denying him an outing at the weekend. If he is perfectly rational he will, according to the Audi, have sufficient motivation to act accordingly provided this is his all-things-considered judgement. If he fails to act accordingly then the weak internalist will accuse him of some sort of practical failure and the strong internalist will argue that he is irrational because of his inferential failure. On both views it is supposed to be evident that unless there is something wrong with him, he must be motivated to deny his son the outing. However it is difficult to accept that this is indeed evident. As Audi points out this moral judgement (that he ought to deny his son the outing) represents only a small portion of his overall structure of relevant beliefs, desires and goals. He might even withdraw his moral judgement if he reflects some more on his son's reasons for taking the car or on his growing need for

[38] The agent who knows, for example, that by walking in the snow she will get her shoes wet and walks in the snow does not *intend* to get her shoes wet – she just intends to walk in the snow.
[39] Audi (1997), p.229-230.

autonomy, for example. The fact that he might actually do this does not imply that he or his moral judgement is irrational while he holds it. Audi's suggestion is that the father is motivationally influenced by the reasons for leniency that he senses but does not draw together into a proper counter-judgement or superseding all-things-considered judgement. It is not the case then that his reason overcomes irrational impulses. His original judgement is *not* shown to be irrational or to be strictly outweighed by any added considerations, rather it is only shown to be less rationally adequate than one he might form from a wider perspective.[40] In this case then it just fails to make sense to claim that it is conceptually necessary that he be motivated to deny his son the outing. The point here is that *contrary motivation does not have to be, by necessity, irrational.* Having an all-things-considered judgement does not mean that all the relevant rational considerations have been exhausted.

Just how broad is the range of the relevant rational considerations? There are not only judgements and facts to be considered, but also moral emotions such as disgust or revulsion (for example, revulsion at seeing your son made unhappy by an unnecessarily harsh punishment), and moral experience, which includes such things as a sense that a given deed would be wrong in some way. Now moral emotion and moral experience or sense are not *judgements* but are considerable forces that can lead to the formation of new judgements and can deprive a judgement of its motivating power. But it is also possible that motivation to act can be blocked by such things as depression, love or pity etc. If this is true then it is not entirely clear that there is any significant deficiency in rationality. But more importantly, what these points serve to show is that weak internalism, or internalism generally, is not a *conceptual* truth and indeed it is to be doubted whether it is even a *psychological* truth. If we try to counter

[40] Ibid., p.230.

this by arguing that the reason why the father failed to punish his son was because he never fully held the judgement that he ought to punish him in the first place (as in Hare's 'no genuine judgement' view) we are in danger of putting forward a trivial and stipulative view of practical reasoning.

Another significant factor here that we ought to consider is that there is a difference between holding a moral proposition – holding a detached judgement with moral content, that is, believing *that* – and judging a matter morally or from the moral point of view – making a move in moral discourse. It could be argued that there is some degree of motivation internal to moral judgements held *from the moral point of view*. If so, then motivation to act on our moral judgements can be internal to holding the moral point of view without being internal to holding those judgements. The rational agent's actions might be motivated by something *extrinsic* such as moral experience or by a desire to do what is right. The suggestion here is that moral motivation may be entailed by, and so be *consequently* internal to, moral judgement held from the moral point of view, even if it is not an intrinsic element of moral judgement by itself.

If this view is correct then there can no longer be claims to the effect that the connection between holding a judgement or having a reason to act and being motivated to so act is logically guaranteed or in any way a conceptual truth. But if it is not a conceptual truth then must we conclude that it is therefore merely a causal truth concerning how agency happens to operate psychologically?

So far we have unearthed several assumptions behind internalism that have given us cause for concern. These can be summarised as follows:

1) That although justifying a judgement does not include giving motivational reasons for acting on it, it presupposes the existence, at least prima facie, of such motivation.

2) That saying or being said to have *a* reason presupposes one's having (at least *prima facie*) motives for doing the action in question.

3) That having a reason entails feeling or having a disposition to feel at least some inclination to act in the way prescribed.

The crucial issue here arises not when we attempt to consider the truth of these claims (on the face of it they seem fairly unobjectionable), but when we try to decide just what status these claims actually have if we deny them the status of broadly conceptual necessity.

To answer this we should ask what sorts of other ways there are in which having a judgement or reason for action can imply motivation to act. The key notion we are dealing with here is idea of *assenting* to what one perceives to be an obligation to act. If we characterise this assent as conceptually necessary, as just *what it is to hold* a judgement or reason, then we find ourselves with the kinds of problems strong internalism has to deal with. If we follow Audi's (or really *Hume's*) suggestion above then we might say that assenting to a judgement entails motivation in the sense that this is just what the rational agent does when considering her prospective actions from the moral point of view. This is a causal claim based upon what we know about the dispositional makeup of the rational agent. But this may only mean that we commonly believe that all agents are psychologically *so constituted* as to typically be moved by the recognition of what they believe they ought to do. We could even combine this consistently with the view that what makes an action obligatory is some *external* characteristic such as an intuition of some right-making fact or property of goodness.

It seems plausible, then, to think that when an agent holds or assents to a reason for action we can expect that she remain committed to performing that action.[41] But it does

[41] The origins of this argument are found in Frankena, 'Obligation and Motivation in Recent Moral Philosophy', in ed. Goodpaster (1976).

not follow from any of this that having a motive to act is a necessary condition on having a reason or holding a judgement about the action in question. Here is what we are left with for weak internalism:

1 That although justifying a normative judgement does not include giving motivational reasons for acting on it, it presupposes the existence, at least potentially, of such motivation.

2 Saying that one has a best judgement in favour of a certain course of action presupposes one's having motives for doing the action in question.

3 That committing to one's best judgement entails feeling or having a disposition to feel at least some inclination to act in the way prescribed.

So it appears that this is what we are left with when we weaken internalism. If the connection between holding a judgement and being motivated to act is characterised in this way then we might think that we can at least be sure that we can successfully explain cases of akratic action that could not be explained by strong internalism, since the latter only allowed for weakness of will to be understood as a problem about inferential failure. If the connection is weakened then many different failures of practical reasoning open up to explain akratic action. For example, if Tom judges that he ought, all-things-considered, to X at t but then does Y instead, we can explain his weakness of will as the result of his not being fully rational. But this is not the same as what strong internalism means by claiming that an akratic agent is less than rational. What the weak internalist is saying here is that Tom is not in a psychological state conducive to acting on his all-things-considered reason for action. This explanation can go further of course. We can perhaps pinpoint the problem – maybe he feels threatened, has no self-confidence, is lazy or lacks a proper sense of the importance of his judgement. Whatever the problem is, the point is that once we see that internalism needs to be weakened and it ends up being construed as a causal and/or perhaps

psychological claim then the door is opened to the vast regions of psychological explanations for weakness of will and other aberrant behaviour. On this view merely to say that an akratic agent is irrational is no longer to offer an answer to the problem that holds any explanatory force, for to label an action as 'irrational' in no way explains what has gone wrong with the agent's practical reasoning. If the connection between having a judgement or reason for action and the motivation to act is a causal or contingent connection then the reason why an agent fails to act as she thinks she should is not primarily about her rationality but must be something to do with weaknesses in virtue in a wider sense or in practical wisdom. Her failure to act then is very much a *practical* failure, not some sort of logical failure analogous to a failure of theoretical reasoning.

However, I do not believe that we should accept that this view is inevitably one that we are pushed towards. While it does seem to be the case that strong internalism of the kind discussed earlier cannot account for all cases of akratic action, we do not, I want to argue, have to accept this weakened form of internalism in order to be able to account for all the different types of action that we believe cannot be plausibly denied. This version of internalism – 'weak internalism weakened' as we could call it – is in danger of dying the death of a thousand qualifications. By taking small, seemingly plausible steps we have arguably arrived at a position that is so weakened that even an externalist could accept it. Externalists can admit that there is such a thing as non-rational motivation and even that any statement that says that having an obligation, assenting to one, or being said to have one, causally or psychologically involves the existence of a corresponding motivation. The externalist might even argue that since we are psychologically constituted in a certain way it is necessarily (causally) the case

that if we have a reason for action, we will (*ceteris paribus*) act on it. We could put this another way: externalists can admit that motivation is *consequently* internal to holding a judgement or reason even if it is not *constitutively* internal.[42] All that externalists deny is that having a normative reason to act implies having a motivational reason to act. For example, consider David Brink's three central externalist claims as he puts them:

1) Moral considerations only contingently motivate or justify.

2) The motivational power or rationality of morality, whether necessary or contingent, can only be known *a posteriori*.

3) The motivational power or the rationality of morality whether necessary or contingent, *a priori* or *a posteriori*, depends on things other than the concept of morality, such as what the content of morality turns out to be, a substantive theory of reasons for action or facts about agents such as their interests or desires.[43]

As we shall soon see below the weakened version of internalism outlined above is not in fact substantially different from externalism. However, one might wonder why this is presented as if it were a consequence to be avoided if at all possible. If we have to weaken internalism what is so wrong with having to accept that externalism might be better able to cope with the kinds of actions that internalism makes conceptually impossible? Well, there are two problems here.

Firstly, how could we blame an agent for acting akratically on the externalist view? If we can always explain the psychological states of an agent who fails to act as she believes she should and these states suggest that despite the agent's efforts her dispositional character is such that it is extremely difficult for her to act on her best reasons for action then other than holding her responsible for her dispositional character itself, how can we hold her responsible for her 'akratic' actions? Externalists explain how it seems unassailable that motivation is reliably connected to having reasons or

---

[42] I take these terms from Robert Audi, (1997) p.226. This point was made much earlier by Frankena, 'Obligation and Motivation in Recent Moral Philosophy' (1958).

[43] Brink, 'Externalist Moral Realism'. With these criteria he thinks the following are externalists: Plato in *Republic,* II-IV, VII-IX, Aristotle *EN* (Irwin 1977:chapter 8, and 1981), Sidgwick 1907:498-503, Frankena 1958, 1973:114-15, Gauthier 1967 and Foot 1972a, 1972b.

best judgements by arguing that the motivational force of moral considerations is a matter of contingent psychological fact. However, this may explain the reliability, but it just does not say anything about *failures* to act. Explaining why a connection is reliable does not give a satisfactory explanation of what is happening when the connection breaks down. All the externalist can really say is that the connection is reliable except when it is not – an occasional akratic breakdown in otherwise normal agents does not amount to evidence of a lack of a (contingent) connection. But the problem with this is that this type of connection is *too* defeasible.

This is the opposite of the problem strong internalism encounters and tries its best to avoid. To recall, the strong internalist of the 'no genuine judgement' variety claims that to hold a reason for action all-things-considered necessarily involves being motivated to act, and if the agent is not so motivated then she does not have the reason to act, or the best judgement, to begin with. This connection is indefeasible to a highly implausible extent as we have seen. By contrast, the externalist's connection by being too *defeasible* effectively claims that if an agent fails to act according to her declared best reason for action all-things-considered then the most we can say is that by not exhibiting the usual reliability the agent may have some malformed disposition affecting her more typical dispositional character. But if agency becomes in this way simply a matter of competing causal factors of varying strength and uncertain operation, then it becomes hard to recognise the general sort of goverance by reasons, however fallible, that we do take to characterise human agency in general. To avoid these consequences (for morality in this case) the externalist might reply that this is prevented by a type of moral realism being true – agents can be blamed for not recognising the objective moral properties of an act. It might even be argued that a

realist does not have to claim that moral requirements provide any agent to whom they apply with reasons for action independent of their desires.[44]

However this brings us to the second problem with externalism – it fails to capture the reliability of the connection between holding a best judgement or reason and being motivated to act. What explains the reliable connection between judgement and motivation for the externalist is the motivational disposition I have in virtue of which I count as a moral person. The content of my motivation in this case must be something like the desire to do the right thing. That is the only possible motivation that could count as a plausible explanation. If I said that I decided to vote for X instead of Y then without the desire to do right there would be no reason for me to be motivated to vote for X. If I had that motivation it would be contingent and optional. It would not be impossible for me then to be in a position where I was not motivated to do what I judged right and also where I was motivated to do what I judged was wrong. Externalism needs a basic motive in order to explain the reliability of the connection between judgement and motivation (e.g. a basic motive to do the right thing). So I would have a judgement that I should vote for X followed by a derivative desire to vote for X because I have a non-derivative desire to do the right thing.

One clear difficulty here is that on this view an agent must always have a desire to do the right thing in any situation where she is to act according to her best judgement. As is familiar[45] this involves stretching beyond plausibility the claim that the moral person is simply one who is motivated to do what is right. But if this is true then even if the externalist admits that there are other non-derivative desires for such things as honesty, justice, the well-being of family etc., these cannot alone motivate an agent to act. So not only must Tom, for example, have a desire to treat his employees

[44] This is Brink's strategy for avoiding the consequences I discuss.
[45] For example, by Michael Smith in *The Moral Problem*, chapter 3.

fairly but in order to do so he must desire to do what is right. Just the former without the latter cannot result in fair treatment. But such a position asks too high a price. What we are being asked to give up is the view (and indeed the commonsense intuition) that morality embraces multiple values like love, friendship, fairness, honesty etc. It is very difficult therefore to see how this kind of externalism could work. The good person has many non-derivative concerns, and in order to work externalism has to give the desire to do right the status of most important or 'one and only' moral virtue. To defend herself here the externalist would need to argue that all the other virtues I mention are derived from our desire to do right plus our beliefs about what right requires, but again it is difficult to see how this could work either. It seems just as plausible to argue that we have a non-derivative desire to do what is just, or to do what is loving as it is to argue that we have a non-derivative desire to do what is right. There just does not seem to be any clear basis on which to figure out what *the* moral motive could be. One cannot help but think that to choose would be purely stipulative.

So it appears then that weakening internalism in order to account for weakness of will can lead us towards an untenable externalist position that not only fails to give any account of the practical (never mind the logical, if there is any) failure of the akratic agent but also fails to capture the reliability of the connection between holding a judgement all-things-considered and the corresponding motivation to act. It seems then, that we are right back at the question of how to weaken internalism in order to account for our problem cases of weakness of will where the agent does appear to act against an already formed intention. If we cannot successfully argue that weakness of will is an essentially *practical failure* do we have to resign ourselves to the view that all we can say is that it is irrationality in the sense of *inferential failure*? I do not believe

so. Further, I intend to argue that we are not forced to choose between inferential failure and practical failure. However what these issues bring quite forcibly to our attention is that the problem of weakness of will cannot be discussed separately from the issue of how best to capture the action-guiding or directive force of reasons for action.

We might say that the weakened internalist position brought us to the 'space of causes', and we can see the problems with this. But there is another way to solve the weakness of will problem that we have not yet discussed. This position, neo-Kantian in influence, ends up making the inferential failure solution to weakness of will conceptually impossible. This is perhaps the best place to discuss this view, for defenders of this position share similar worries about the motivational internalism we have been discussing until now. Jean Hampton, for example, agrees that the only form of motivational internalism that is plausible is a very weak form that is problematic for many of the same reasons we have pointed out.[46] Hampton and others suggest that we abandon humean-influenced attempts to capture the directive force of reasons for action. Reasons for action are action-guiding not because they are necessarily linked with motivation, but rather because there is a *normatively* necessary connection between judgement and motivation which has its source in the normative authority of reason. We are not talking here of any kind of *causally* necessary connection. The normativity of reasons sets conditions on having a reason to act. Desire is just one among the antecedent conditions of a rational motivation and as such the influence of desire can only be seen when we examine how it interacts with the normative structure which explains how the influence of desires is extended through a rational system. There is a difference, on this view, between the *authority* and the *motivational efficacy*

[46] Jean Hampton, *The Authority of Reason* (Cambridge, 1998), see especially chapters 3-4.

of reasons, and such views as motivational internalism have been put forward because these two factors have been confused in the attempt to naturalise reasons. Weakness of will *is* irrational on this view because it goes against the normative authority of reason.

Having established that the problem of weakness of will cannot be treated in isolation from questions about how to characterise the reliability of the connection between judgement and motivation I will just briefly point in the direction I believe we will find our solution. In the next chapter I intend to argue that the neo-Kantian view just mentioned depends on a sharp and untenable distinction between the normative and the descriptive as well as on a mistaken view of deliberation. This is no easy task, for clearly there are certain phenomena of moral experience that this neo-Kantian view insightfully captures. One of these insights concerns what we experience as the *inescapability* of ethics. When we assent to a moral judgement we no longer feel that we can just take it or leave it – we cannot avoid *being obligated* by our acceptance of a course of action as best, all-things-considered. But if we appeal ultimately to desires to ground our sense of obligation then it becomes much more difficult to see how any agent can be obligated to act independently of a desire to do so, and so the very inescapability considered central to moral experience on such a view suffers too-severe qualifications. Relevant here is another insight concerning the apparent *objectivity* of moral judgement.

Despite the complexities involved in the moral life, we do suppose that there are answers to moral questions, that there are in some sense moral facts determined by circumstances and that our moral judgements express our beliefs about what the facts are. We commonly believe that moral norms apply to us 'objectively', that is, to say that we have reasons for action is to say that these reasons are not merely invented by

us to safeguard our own or our community's interests or concerns but are considerations "outside" ourselves to which we are responding when we believe we are obligated to act.[47] Furthermore, we can still recognise that reasons have moral authority even when we have no motivation to act accordingly and this important intuition seems difficult to account for on any view which depends on agents having independent basic desires in order for them to have even a prima facie reason for action. In so far as this perspective is successful its adherents believe they can keep at bay all those views that have a tendency to weaken the inescapability of the force of moral reasons such as error theories and expressivist theories. But there is a limitation here that even Kant found deeply perplexing:

> ...[I]t is quite beyond the capacity of any human reason to explain *how* pure reason, without other incentives that might be taken from elsewhere, can be of itself practical, that is, how the mere *principle of the universal validity of all its maxims as laws* (which would admittedly be the form of a pure practical reason), without any matter (object) of the will in which one could take some interest in advance, can of itself furnish an incentive and produce an interest that would be called purely *moral*; it is impossible for us to explain, in other words, *how pure reason can be practical,* and all the pains and labor of seeking an explanation of it are lost. (*Groundwork*: 4:461)

The difficulty Kant finds himself with here concerns the extent to which non-instrumental or 'pure' reason in itself can be motivational. If motives for being moral cannot be derived from anywhere else other than ethical principles themselves then there is no sense in which the force of ethics depends on anything like a basic and ethically-neutral desire for X. So our ethical principles would govern us just as Hobbes and Hume believed that prior motivations do. But then there is the closely related difficulty: what grounds or validates our ethical principles themselves? Kant in effect argues that it is our rational recognition or endorsement of them as authoritative that does so; but this just serves to reinforce the need for some account of how pure reason, morality, and moral motivation are really linked up. Ultimately then for Kant, in order

[47] See Hampton (1998), Nagel (1970), Korsgaard (1986) for good examples of the strength of this view.

to explain how it is that practical reasons can motivate an agent to act we need to rely on a conception of ourselves as free, but all we know is *that* this sort of freedom is necessary for morality and possible in the sense of thinkable without contradiction (4:461). What remains intractable is *how* to justify or ground this presupposition. Not surprisingly these issues concerning 'pure practical reason' and 'transcendental freedom' and treated gingerly or often re-conceived by contemporary neo-Kantians.

The questions we have to ask next are: (a) to what extent have those holding neo-Kantian views succeeded in showing us that something is not a reason simply because it motivates, but that reasons are nonetheless inescapably motivational? (b) To what extent can neo-Humeans satisfactorily present us with surrogate explanations for the insights captured so well (but ultimately problematically) by neo-Kantians? The nature of the solution or, more precisely, an explanation for weakness of will depends on which of these perspectives succeeds, and correlatively it is the problem of weakness of will that serves to point a necessary spotlight on the nature true of the directive force of reasons.

# Part II

## Weakness of will and Scepticism about Practical Reason

One clear conclusion that has emerged from the previous chapters is that the problem of akrasia cannot be engaged, never mind solved, independently of the question regarding how to characterise the reliability of the connection between judgement and motivation to act. As we will see in chapter three, something like this conclusion is accepted by, for example, Christine Korsgaard, who argues that scepticism about the motivational capacity of practical reason can only be based on scepticism about the capacity of practical reason to determine what is valuable. If practical reason can determine what is of value then, we are told, there is no justification for doubting whether practical reason can motivate agents to act since every rational agent would accept that they are committed to doing what they themselves consider to be rational. On this neo-Kantian view weakness of will and other irrational actions are simply examples of conditions that can cause us to act irrationally, but they cannot be used to deny that there must be principles of conduct other than hypothetical imperatives based on desires.

In fact weakness of will cannot even be explained on the opposing view of reason as instrumental, or so Korsgaard's argument goes. For if practical reason is merely instrumental, if agents only have reason to do what they have most motive to do, then the agent who fails to act on her greatest motive and acts on a lesser one can no longer be said to have had the greater motive in the first place and so, rather ironically, weak-willed actions cannot be accounted for. This is supposed to further strengthen the case for a non-instrumental view of practical reason, but I will argue that Korsgaard and her fellow neo-Kantians are wrong here and that it is not true that the kind of problems about weak-willed action we have been dealing with are merely motivational

ones that can be psychologically explained. This will emerge as a question-begging response to the problem of weakness of will.

Here (in Part II chapter three) we will examine in some detail the objections that lead to such a response before we go on to chapter four, where I will contend that those favouring the neo-Kantian or Pure Practical Reason view end up begging the question regarding weakness of will because of their view that one endorses a reason for action only when one evaluatively concludes that that reason is best, all-things-considered. This assumption is one I believe we should reject and replace with a different understanding of what it is to endorse a reason as one upon which the agent should act. In Part III I attempt, in outline, to tell a somewhat different story about the relation between endorsing a reason and acting on it. However before we jump too far ahead of ourselves, let us briefly overview the issues to be addressed in Part II.

Crudely speaking, if we accept the view that we do not find value *in* the world but instead *construct* it from our own reasoning about how to act, then it seems that we ourselves are entirely responsible for the *content* of morality or the moral principles we hold supreme. If this is to be a correct (even if very broad) characterisation of morality then it is obvious that our practical reasoning or reasoning about how to act is our qualification (as it were) for the difficult business of figuring out the right way to act in all situations where choice of action is called for – and that is just about all the time for most of us. It is also obvious that practical reason needs some principles according to which we can determine how to act. In other words, practical reason requires *content*. It is here however that views diverge. Just what is to be the content of practical reason? According to one view, if reasoning is only concerned with the abstract relations of ideas or with matters of fact and causal relations then it cannot tell us what our goals ought to be; and if it cannot tell us this then we must look elsewhere to discover the sorts of things that might be the worthy ends of our actions. Desires or predispositions

usually fill this role in the sentimentalist tradition, most notably represented by Hume. Reason can only tell us *how* to achieve our ends not what those ends are to be. On this view an agent's reasons for engaging in a particular course of action cannot motivate her to perform the act in question. The act to be performed must be, or be the means to, a desired end of the agent's if she is to stand any chance of really performing the action. So in order for some consideration to even be a reason for action the agent must have its object as a desired end. This view, we are told by its opponents, amounts to scepticism about the motivational power of practical reason; it is based entirely on a view of practical reason as solely instrumental – or to put it another way, on a general scepticism about the possibility of practical reason having any substantive content.[1]

On such a view[2] it is easy to see the significance that irrational action in general is supposed to have. A weak-willed agent commits herself to acting according to what she takes to be the best or most appropriate course of action and yet despite this she fails to act. The explanation given is that she is not *motivated* to act. There are a variety of explanations for why this occurs, as we have seen, but the important point here is that she is just not motivated to act, there is something missing and that must be the *desire* to act. But the weak-willed agent is not too far removed from the rational agent. The difference between them is that the rational agent's desires are in line with her reasoning about how to act whereas the weak-willed agent's are not. If certain types of agents – irrational agents generally – can have practical reasons in favour of acting and yet fail to act then there must be something wrong with the view that practical reasons are sufficiently motivational.

Like Korsgaard, Nagel and Hampton among others I do not believe that this is a satisfactory account of the significance of weak-willed or irrational action. While I do

---

[1] See Korsgaard (1986).

[2] The 'IR' view for short. A rather crude Humean view but nevertheless this seems to be a common understanding of 'Humeanism'.

not accept their view that it is not possible to place independent constraints based "solely on motivational considerations" on "what might count as a principle of practical reason,"[3] the main purpose of this chapter is *not* to argue that Hume or Humeans do attempt to, and succeed in, deriving scepticism about the capacity of practical reason to have substantive content from scepticism about the capacity of practical reasons to motivate. Rather, it is to show that a careful examination of the problem of weakness of will makes it clear that arguments against desire-based accounts of practical reason on the basis that they *cannot* account for irrational action because they conceive of rational agency incorrectly are problematic and do not give us sufficient reason to reject a broadly Humean account of practical reason. Korsgaard, for example argues that only a view of practical reason as non-instrumental can account for weak-willed or irrational action at all. The position Korsgaard holds appears to be consistent with the conclusion we arrived at in the previous chapter, that is, that an examination of the problem of weakness of will shows that the most familiar versions of motivational internalism are insufficient as an explanation of what it is to act on, or have, a reason. Korsgaard however, and other Kantians similarly positioned, see this as a further strengthening of their case for a theory of practical reason as having "substantive content."[4] Their argument is that a Humean account of practical reason (the 'instrumental reason' view) cannot account for weakness of will because the conception of rational agency it advocates is wrong. An IR view requires prior normative commitments in order to be able to explain how actions are evaluated or how agents can resolve conflicts of ends. The problem, according to the Kantians, is that an IR view cannot have any concept of a *normative* commitment since it is committed to the instrumental principle as the sole principle of practical reason, and this principle

[3] Korsgaard (1986), p.8.
[4] ibid., p.5.

just cannot provide the normativity required. If such a view cannot account for problem cases of practical reasoning like weakness of will, then it is going to have similar difficulties explaining all evaluative practical reasoning generally. In other words, it fails completely as an account of practical reason; all an IR view can really do is *describe* how agents do in fact act.

However I do not believe that this argument against the very possibility of an IR view (and not just against particular IR views) succeeds in showing us either that all IR views are doomed to merely describe agents' acts or that they do not have the conceptual tools to explain weakness of will and the evaluation of reasons. The purpose of Part II is to discuss the main objection to the possibility of an IR view. This objection centres on the problem of the normativity of any hypothetical imperative and can be described briefly as follows. The instrumental principle cannot stand alone in grounding a theory of practical reason because there are or must be structural constraints on practical reason already in place that have a crucial influence *before* an agent reasons instrumentally. Without such norms it is claimed that the following cannot be explained: (a) the ranking of ends or goals; (b) preference change over time; (c) weakness of will.

However before we begin our discussion of these objections, let us take a few moments to put things into a broader perspective. It is very common in the contemporary philosophical scene to find a great deal of doubt about the Kantian conception of 'pure practical reason' (the 'PPR' view). The objections raised concern the nature of the principles of pure practical reason that a Kantian could come up with and the fear is that these principles would be so constrained as to be purely formal – an empty proceduralism with no substantive content. There is heated disagreement, for instance, about the capability of broadly logical requirements such as consistency to settle practical questions. Now I do not want to pursue these issues here; we cannot

settle here the question of whether Kantians succeed in defending categorical requirements on practical reason. But there is another way open to Kantians to defend their views (exploited to some extent by both Korsgaard and Hampton) and that is by attempting to show that despite all the worries about a Kantian view of practical reason, the alternative views are in an even more troubled state. The main route available for this purpose corresponds with the chief objection outlined above. If this and related objections to an IR view are found to be sound then it could be argued that a PPR view has more promise despite worries about purely formal principles etc.

However I am troubled by this strategy. For if the central objection to the IR view turns out *not* to be sound then whatever we think about the Kantian project, it cannot defend itself via this strategy. The problem is that some Kantians (such as Korsgaard) seem to think that with these objections they are rejecting the very possibility that an IR view could provide a working account of practical reason. However what I hope will eventually be clear is that these objections will not work against a sophisticated IR view of practical reason. (In fact, I do not think they entirely work against even Hume's own view but I will not attempt to defend this here either.) Rejecting one form of substantivalism does not establish the validity of proceduralism. Or, at risk of sounding crude, rejecting Hume does not get us Kant. Coming up with a satisfactory account of weakness of will is central here since if the charge of descriptivism can be defeated then we can begin to address the issue of the motivational and normative status of all-things-considered best reasons for action.[5]

[5] That is, we cannot address this issue until we can make a distinction between what an agent's end or goal is and what she thinks she should do (all-things-considered).

# Chapter Three

## Pure Practical Reason and the 'True Irrationality' View

### 3.1 Motivational Scepticism

One of the most common steps in the attempt to show that pure practical reason, if it exists, has no real opposition to face from other theories of practical reason, is the argument that seeks to demonstrate that motivational considerations provide no grounds for doubt about the attempt to base our view of what it means to have a reason for action on *reasons internalism*.[1] The suggestion is that those who put forward motivational considerations as sufficient reason to doubt the directly motivational nature of an all-things-considered reason for action are confusing the claim:

(1) that all-things-considered reasons are necessarily motivational;

with the claim:

(2) that an all-things-considered reason never fails to motivate.

It is obviously false to hold this latter view, and if this is what pure practical reason amounts to, then we have no good reason to accept it. Korsgaard, for example, maintains that this confusion is based on a misunderstanding of what is required by internalism to begin with. Reasons internalism does not require that any agent who accepts that she has an all-things-considered or overriding reason for action necessarily must act or else we are entitled to conclude that she must never have had the reason she took herself to have in the first place. There are serious difficulties with such a view as we saw above in our discussion of the 'No Genuine Judgement' view.[2] To this extent we can agree with Korsgaard. However we can, I believe, question the conclusion she

[1]Reasons internalism is the view that at least some reasons for action are directly motivating. This is analogous to, for example, what Kant sees as the immediate and direct motivational efficacy of a categorical imperative.

[2] See Chapter 2, section 2.3.

herself reaches, namely, that when an agent fails to act as she believes and accepts she should what we are seeing is nothing more than the limits of that particular agent's rationality.

It is only in so far as an agent is rational, we are encouraged to believe, that she will necessarily act according to her all-things-considered reasons for action. But since we are not concerned with the extent to which people are *in fact* rational the task of figuring out how we are to understand practical reason need go no further than explaining how it is we act on our reasons *in so far as* we are rational. The most important step in this argument is the claim that reasons must be capable of motivating us. Thomas Nagel, for one, makes this point clear: in order to fully explain an agent's actions, reasons must be motivational. We must say then that it is a requirement of practical reasons that they motivate since otherwise they just could not explain how it is that we act for reasons. This is another formulation of the internalism requirement we have seen in operation earlier.[3]

Now the argument I want to bring into question claims that the internalism requirement has no capacity to impose any limitations or restrictions on what might count as a principle of practical reason.[4] On a view such as Korsgaard's, it would seem that the internalism requirement together with the problem of weakness of will cannot combine to allow us conclude that scepticism about practical reason is justifiable. However I want to reject this argument and contend that the internalism requirement together with weakness of will do entail that one ought to be justifiably sceptical about the claims of pure practical reason.

Let us take a look at the former argument first of all. Scepticism about practical

---

[3] This can also be expressed as the following: 'If an agent thinks she has reason to Q, then *ceteris paribus*, she will Q.' See also Nagel (1970) chapter 2 for a similar expression of this motivational claim.
[4] For example, See Korsgaard (1986) p.23.

reason concerns the "priority of psychology."[5] This has held theories of practical reason hostage far too long, so the story goes, and instead when we investigate and come up with the right moral principles *these* will be reinstated to their rightful place as fully able to set the necessary rational conditions on desire and action. In other words, principles of practical reason will have serious implications for the type of motivational psychology the rational agent is conceived to have.[6] What these principles of practical reason are is a matter of debate among those who adopt this strategy but the essential point we are concerned with remains the same.[7] Reasons in order to be reasons must motivate if they are to have any role in the explanation of action at all. If we can show that there really are such things as reasons for action then we will at the same time have shown that there are some things capable of motivating us. If in turn these reasons are derived from ethical principles then we can count these principles among the basic elements of motivation theory. As Nagel tells us, ethical principles cannot be derived from our motivations or predispositions, instead they "define motivational possibilities rather than presuppose them." So ethical principles can govern us just as Hobbes and Hume think that our prior motivations govern us.[8]

But those holding this view can find themselves dealing with the same difficulties Kant originally found himself with, in particular that it has to be *independently shown* that we can be motivated to act on rational principles (or for Kant, according to the categorical imperative). Pure practical reason *can* and *often does* motivate us to act, but Kant stops short of arguing that free rational agents *necessarily* act morally and maintains instead that they *can* act morally in so far as they are rational. The internalism requirement is only supposed to be interpreted as an

---

[5] Korsgaard (1986) following Nagel (1970).
[6] Nagel (1970), pp.13-14.
[7] We can either look to Plato and Aristotle for examples of non-ethically-neutral motivation or to the more common modern source of this view, Kant.
[8] Nagel (1970), p.14.

expression of how in fact our motivation works most or at least some of the time, at least in principle. If we fail to act as we think we should then there is a problem with our rationality only in the sense that we have failed to be fully rational, and not in the sense that we are not beings for whom pure reason is practical. The system may not be working correctly but the system remains in place nevertheless.

In Kantian terms the opposing view holds that it is perfectly possible to have an autonomous will that operates with the hypothetical imperative as its sole rational principle of action. In other words, it is plausible to conceive of practical reason as working on the principle that it is rational to take the necessary and permissible means to one's ends or else abandon those ends. This does not necessarily mean that the ends in question are only pursued or willed because they are *desired*; we can remain neutral on that point for the time being at least. Now opponents of the instrumental view argue that in fact the PPR view is the inevitable result of thinking through the problems inherent in this opposing IR view. That is, if we start out with the internalism requirement, and with the belief that this, along with motivational scepticism, can place constraints on what could be a plausible theory about the content of practical reason, then it is argued that we really have no plausible option but to recognise the necessity of the PPR view. This is indeed the heart of the matter for us when it comes to considering both the explanation and the significance of the problem of weakness of will, since figuring out how to understand irrational action plays a central role in both views. However in order to see the differing treatment of weakness of will let us first examine the challenge involving the explanation of irrational action proponents of the PPR view pose to the so-called motivational sceptics. If this challenge can be met satisfactorily then the PPR view will be much less the inevitable position it is argued to be.

One commonly-held view of those holding the Pure Practical Reason or PPR view is that the Instrumental Reason or IR cannot account for irrational action because it cannot give *normativity* to the instrumental principle.[9] In other words, it fails to give a satisfactory account of how this principle is supposed to govern action in the sense that it ought to do if it is a rational requirement of agency. There is one central theme that tends to dominate objections to the IR view. Put broadly, it centres around the charge of descriptivism and is something like the following:

> No view of practical reason that holds the instrumental principle (or indeed any hypothetical imperative) as the sole normative principle governing action can explain the various forms of irrationality that occur; rather it threatens to make weakness of will impossible to account for.[10]

What I hope to show over the course of this chapter and the next is that neither this nor any derivative objections are sufficient to allow us conclude that weakness of will is impossible on an IR view of practical reason. For the remainder of this chapter we will be dealing with issues arising from this main objection.

### 3.2 The Instrumental Principle: Norm or Description?

Proponents of the PPR view argue that the instrumental principle cannot stand by itself as any sort of an authoritative norm governing action. The problem with it runs as follows. If we commit ourselves to the existence of the instrumental principle as the sole norm of rational action then we must believe that reason, in the practical arena, is

---

[9] As mentioned already it is not necessary that this IR view specify that the ends of any agent be willed or intended because they are desired by the agent or are part of her dispositional set. There is room for interpreting the instrumental principle as a non-desire-based rational principle.

[10] For example, Korsgaard (1986; 1996); Hampton (1998); Darwall (1983); Smith (1994); Pettit and Smith (1993).

solely about figuring out the means to already determined ends – ends which are chosen independently of any operation of reason other than adequate object-representation, or determining causes and effects or some such. So action occurs at the prompting of what Hume calls 'passion' or what we would call our predispositions. Hume argues that we have what he considers "a general appetite to good, and aversion to evil"[11] (where good and evil for Hume are, "in other words, pain and pleasure"[12]) and this appetite is in the category of our 'calm passions'. Insofar as this appetite remains predominant agents will seek to achieve their chosen ends through following the dictates of their instrumental reasoning. If an agent must choose between *different* and *incompatible* ends then in light of her general appetite to the good she will most likely act to achieve the end most in line with her anticipated good. But Hume's famous proviso here lands us in more than a little confusion:

> 'tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter.[13]

This often appears to be nothing more than a colourful statement of Hume's avowed position on the purely instrumental nature of practical reason. However when we consider this along with the claim that we have a general desire for the good then it can become difficult to see how Hume could consistently allow for irrational action at all, and this is what PPR defenders focus on. If the general desire or appetite for the good lessens then there will be no motive and so no reason for doing whatever it is that will bring about one's greater good.

For example, John has a general desire to seek what he considers to be in his best interests – a job in the civil service. Now when he thinks it through he realises that

---

[11] *Treatise,* II.iii.3, p.269.
[12] *Treatise,* II.iii.9
[13] *Treatise*, II.iii. 3, p.267.

he will have to pass the foreign language proficiency test if he is to have a chance at the job, but since he finds learning languages difficult he also realises that he will have to take an intensive course in preparation. So we can say that John has a motive to act – he wants to reach his goal and get the desired job – and in this case his motive is in line with his reason to act. But let us imagine John wakes up on the morning of the test confident of passing but no longer wanting the job because he no longer seems to have a general desire for his own good. In fact, his aversion to what will do him harm has eased and he now wants to give heroin a go instead.[14] Hume is somewhat ambiguous about whether it is possible to act contrary to the instrumental principle at all. For example he claims that when we "...choose means insufficient for the design'd end" it is only because we "...deceive ourselves in judgement of causes and effects." But as soon as we realise our error "...our passions yield to reason without any opposition."[15]

However the result of this is that Hume looks more like the early Plato of the *Protagoras* as it amounts to the claim that no one ever acts irrationally relative to their own beliefs. Indeed his claim above that it is not irrational to prefer one's lesser good to one's greater appears to lend support to this picture. But without trying to stitch together what Hume may have accidentally left undone, he does appear to intend to leave open the possibility of apparently irrational action. Yet if the impossibility of accommodating irrationality is not the consequence of his view then it appears to be the logical conclusion of a *humean* view at least. It may not be 'contrary to reason' to prefer to scratch my finger rather than achieve world peace, given my predispositions and stated goals in life, but if practical reason is instrumental then I am, as a rational agent, committed to taking the means to my *preferred* end. If I do not do so, then provided I am not *incapable* of taking the means then I can be accused of irrationality

[14] This is not to be confused with the case where an agent is tempted by immediate gratification to forgo long-term best interests. Let us suppose then that John knows all about the dangers of heroin and believes that once he starts there will be no stopping that awful spiral downwards.
[15] *Treatise*, II.iii.3, p.267.

relative to my own judgement of what I should do *all-things-considered.*[16] On a humean view it does seem to make more sense to claim that the instrumental principle is a norm a rational agent follows and provides a standard against which an agent's actions can be judged even if only relative to her own beliefs. In Hume's defence we can say that while it may not be irrational to *prefer* one's acknowledged lesser good to one's greater it *is* irrational not to take the means to an end one wants most of all (all things being equal). But contrary to Hume's doubts about the matter, it seems right to point out that we *can* conceive of an agent who does not make mistakes about relations of cause and effect yet who nevertheless just fails to take the means to her desired end: the weak-willed agent.

However Hume's view is troublesome in another sense. If an agent fails to take the means to some end she has then according to Hume the agent will, as soon as she realises that she has made a mistake about the sufficiency of the means, amend her actions accordingly and engage in the proper means to the end. The problem is that if the agent has not made any mistake about the means and she still fails to act then it does seem as if we must conclude that she never really had that end in the first place.[17] Hume then seems to be saying that if an agent fails to do what she has determined she will do then she must all along have had a stronger predisposition in favour of some other action and therefore really only had reason to do this other action. Perhaps then we could accuse her of self-deception, that is, of not knowing her actual ends or goals.

Proponents of the PPR view such as Korsgaard, Nagel, Hampton, et al., have pointed to this difficulty in Hume's view and tend to interpret it as an insurmountable

---

[16] This is important here since we cannot call the agent who simply has a *conflict* of desires irrational. If Mary has two conflicting desires and in a somewhat bewildered frame of mind fails to take the means to achieve one of them rather than the other then she is not irrational but merely acts too soon – before she has prioritised her desires in light of her best interests.

[17] This of course amounts to the 'no genuine judgement' view we have examined earlier.

obstacle to the possibility of basing a theory of practical reason on anything like an instrumental norm solely. Korsgaard for example claims that

> if true irrationality is excluded and you fail to take the means to some end, this is evidence either that you don't really have this end or that it is not the most important thing to you.[18]

What this means in practice is that an agent like John above, who fails to sit the language test either does not care about his best interests or he cares more about his goal of trying heroin. This is the interpretation of cases like John's we would be forced to uphold, so Korsgaard believes, on Hume's view. But this is clearly counter-intuitive for it amounts to the claim that Hume cannot explain, and does not allow for, the possibility of weakness of will and irrational action generally. If John goes ahead and tries heroin then we have to say that he never really had any reason to do anything else in the first place and so he is not weak-willed. If the action an agent performs is the only clue to that agent's real reason for action then it is not possible to account for weakness of the will. This is because we have to say that the action the agent performed is the action she really had as her end in the first place. In other words, there is no capacity on this view for arguing that the agent in fact acted on a lesser reason for action, one which she already considered and overruled. And so, if weakness of will is acting on one's acknowledged lesser reason for action then this IR view, according to its critics, will never be able to accuse any agent of being weak-willed. This amounts to a denial of the second of our key intuitions about action[19] – that agents sometimes fail to act as they think they should. If it is not possible to account for weakness of will then it would seem that IR defenders are forced to the conclusion that weakness of will does not exist or that it was never anything more than a pseudo-problem waiting to be dissolved. Korsgaard objects to Hume that his view

---

[18] Korsgaard (1986) p.16. What she means by "true irrationality" here is "...such phenomena as self-deception, rationalization, and the various forms of weakness of will." (ibid.)
[19] See 'Introduction', pp.3-4.

> ...makes it seem as if your greater good is an end you might care about or not, and rationality is relative to what you might care about. But once we admit that one might from some other cause fail to be responsive to a rational consideration, there is no special reason to accept this [Humean] analysis of [weakness of will].[20]

As mentioned above, Korsgaard argues that this Humean motivational analysis that is supposed to be equivalent to what others would consider to be a piece of practical *reasoning* cannot impose any kind of restrictions whatsoever on what is to count as a principle of practical reasoning. An agent might not be motivated to take the means to an end that will further her best interests. This fact alone "...cannot by itself throw any doubt on the argument, whatever it is, that preferring the greater good is rational."[21] What Korsgaard is attempting to do here is twofold: (1) to show that the instrumental principle cannot act as the sole principle of practical reason because it is limited; and (2) to argue that no motivational analysis of a situation where an agent fails to act as she thinks she should can impose any conditions or restrictions on the internalism requirement. Her point is that how we explain the motivational elements of any such situation depends upon what we take to be the content (or lack thereof) of rational principles of action. For example, if I think that reason is a slave of the passions and has no substantive evaluative content then I might say that the ends of action are determined by preferences alone, and it is only after that that I can try to argue that weakness of will is impossible.

Korsgaard is not alone in this view. Jean Hampton, another defender of PPR, argues that an adequate theory of practical reason is one that preserves the idea that we can act rationally *and irrationally* and must defend the idea that we can and should act *from* and not merely *according to* reasons.[22] [23] She argues that only rational agents can

---

[20] ibid.
[21] ibid., p.17.
[22] Hampton (1998), pp.160-2
[23] Hampton (1998), pp.142-50. Hampton argues that a naturalist could not consistently hold the view that instrumental reasons have objective authority. Naturalists object to the idea that moral, or, more

be said to *will* ends as opposed to merely wanting them.[24] What she has in mind here is something like a psychological state of *commitment* to chosen ends. However she argues that no humean could admit of such a state since they "...could not say that if this commitment failed – so that this psychological state somehow changed – he had in any way made a mistake." She goes on to point out that "[t]he charge that such a person made a mistake relies on a norm dictating the persistence of this state until the end is achieved."[25] The point Korsgaard makes is being made again here in Hampton's work. The weak-willed agent fails to do what she thinks is best, but Hampton's point is that the agent (as conceived by the Humeans) in this situation might want the end in question but she does not *will* it, and we can rightly suppose that if she fails to will the end then she does not really want that end most of all or does not really think it is best all-things-considered. Therefore irrational action is not a real possibility.

Thomas Nagel finds the Humean picture similarly problematic. He claims that if we argue that ethical assertions are really expressions of inclinations of some sort and then maintain that we always act (i.e. have *reason* to act) on those inclinations, no obligation will be binding on an agent who does not have the inclination to act. So long as ethics is thought of as being derived from predispositions that are immune from any

---

broadly, normative claims generate norms that have some sort of objective authority. On a naturalist view of reasons moral properties are supposed to be understood as being identical with or constituted by natural properties because they supervene on natural properties. This is as close to moral objectivity as naturalists are commonly taken to get and Hampton's view of naturalism is a good example here. Presented like this the objection is fairly obvious – there is no room on this view for the idea that morality is about reasons for action and at best all the naturalist can do is talk about agent-relative reasons. Hampton wants to make a case for moral objectivism by joining a growing number of philosophers such as Hilary Putnam in trying to show that science itself is a normative enterprise. If it can be shown that there is no branch of philosophical theorising that does not depend on a prior theory of normativity then many of the debates about moral judgements will be obsolete because they will have been shown to rely on false or primitive distinctions between such things as facts and values. However Hampton and others holding similar views fail to see any plausibility in the view that naturalism can accept that reason-generating norms of rationality can play a constitutive role in the mental states of agents engaged in ethical situations. If we can give a working outline of such a naturalist view then the way lies open for an instrumental account of practical reason, or we might find at least that such a view does not find its logical conclusion in something like Korsgaard's neo-Kantian ethics of autonomy. We will be dealing with this issue in detail in the next chapter.

[24] Where willing is a distinctive psychological state that also involves reasons as a motive for action.

[25] Ibid., p.165.

rational evaluation we will continue to have these implausible conclusions.[26] If we argue that there is a component in human motivation which can be understood independently of ethics but on which the force of ethics ultimately depends then we will never be able to see situations of motivational conflict as anything other than a clash of competing basic desires or predispositions. The IR view, according to Nagel, will never be able to do anything more than see John's wish to join the civil service as clashing with his desire to try heroin. On this view since one always acts on one's strongest desire weakness of will will be impossible.

The only possible way out of this for the IR view, Nagel tells us, is ultimately unworkable.[27] The humean could say that it is necessary that someone who is fully rational will adhere to requirements of practical reason. But Nagel thinks that this just cannot give us the necessity or authority of ethics he thinks is essential (*a priori* even) because rationality *just is* adhering to rational requirements. In other words, if rationality is adhering to rational requirements we cannot figure out what these are by trying to define rationality. To try to do this would be a useless and empty enterprise because all we could say of a rational agent is that she will necessarily adhere to rational principles. So whatever these requirements of rationality are they cannot be derived from the concept of rationality, they must be established independently. If we tried this strategy with theoretical reason we would see the empty formalism that results quite starkly. To say that requirements on theoretical reason *just are* defined by how a fully rational being would necessarily think is the same as saying that how a fully rational being thinks is how a fully rational being necessarily thinks.

Nagel's proposed solution and alternative to this problem, however, comes from what he sees as a normative view of both theoretical and practical reason. He argues

---

[26] Nagel (1970), p.4.
[27] Ibid., chapter IV.

that neither in the practical case nor in the theoretical case can we analyse rational requirements in terms of simple entailment (presumably from non-rationally informed desires in the practical case and from what the ideal theoretical inquirer would hold in the theoretical). We must instead accept that there must be principles which govern how reasons get to motivate actions. Placing a desire between a reason and an action will not help the humean solve our problems. We still need a framework of some kind with "...principles by which beliefs and desires combine to yield justified actions."[28] Curiously Nagel argues that we need something like Davidson's 'principle of continence'[29] or something that enables us to correct ourselves and accept criticism. However Nagel thinks that this requirement is not a *necessary consequence* of some further principle because then *that* principle would be more fundamental or absolute thus leading us into an infinite regress. Rather the modal status of this principle must be somewhat similar to that along the lines of a non-deductive inference. So the whole basis of practical reason is normative and not "parasitic" on some kind of psychological necessity. Rational requirements then have a sort of natural or psychological necessity but not a categorical necessity and they "...apply to us in virtue of particular deep features of our make-up, features which we cannot alter."[30]

---

[28] Ibid., p.21.

[29] Davidson understands this 'principle of continence' to be: if the agent *believes* that all the relevant reasons support a certain course of action then she should pursue that course of action. Now this principle is not part of the logic of practical reasoning and acts more like an assumption that the agent will act rationally – except when she does not. What this principle does is charge agents with a degree of consistency. We cannot assign beliefs, preferences, intentions etc. to a person in a piecemeal fashion because we can only make sense of such propositional attitudes in terms of other such attitudes. As regards interpreting our weak-willed agent, we could not even begin to have any idea where she goes wrong unless we first attribute to her the same degree of consistency or reasonableness we attribute to ourselves and any other rational agents. Just as the *requirement of total evidence* for inductive reasoning (which suggests that we should accept the hypothesis that is supported by all the available relevant evidence) argued for by (to take one example) Hempel (*Aspects of Scientific Explanation,* pp.397-403) and is not considered to be part of the logic of inductive reasoning, Davidson argues that the principle of continence is analogous and is a principle which rational agents *do* follow despite the fact that it is not a necessary condition of practical reasoning. Davidson (1980), p.41; also see 'Mental Events', (1980), p.221.

[30] Nagel (1970), p.22.

Now what is ironic about this position as put forward by Nagel is that the principles he is looking for, these rational requirements, are offered to us in different guises by the majority of contemporary philosophers who would consider themselves as holding views that Nagel is opposed to. Davidson for example, clearly recognises that any theory of practical reason must be able to explain rather than *explain away* irrational actions and he accepts that it is wholly inadequate to find ourselves with a theory that permits no action to be considered irrational even from the agent's own point of view. As will become apparent those who hold the IR view do see the need for a *normative* requirement to provide the authority or guidance agents need in order to make decisions about how to act.

But if the proponents of the PPR view and those of the IR view both see such a need why then is there such a chasm between them? The most significant difficulty here concerns the modal status of the instrumental principle. IR proponents argue that there is a pivotal sense in which this principle can play the normative role which seems so necessary for any theory of practical reason. However proponents of the PPR view deny that the instrumental principle can ever by itself be sufficiently normative to do the work of grounding a theory of practical reason. Their criticism tends to take the following form: instrumental reasoning cannot stand alone because there are and *must* be many structural constraints on practical reason already in place that have a crucial influence *before* an agent reasons instrumentally.[31] The fatal flaw that supposedly affects instrumental reasoning is that if it is to ground a theory of practical reason it must have some conception of the ideally rational agent's set of coherent preferences or ends (i.e. the ranking or prioritising of ends), but this is precisely what it cannot generate.

---

[31] Korsgaard (1996, 1997), Hampton (1998).

There must be some way of figuring out how to resolve a conflict among my ends. But if our ends are unmotivated (that is, not derived from considerations we might have)[32] and if reasoning about how to act is just a matter of figuring out the best means to our chosen ends, then it does not seem likely that we can resolve a conflict of ends at all. Of course if we cannot resolve such a conflict then it is difficult to see how action could get started. On traditional readings of Hume's (IR) view reason plays no role in determining an agent's unmotivated ends so there is no evaluation of the agent's preferences or what she declares is her good or in her best interests. The argument from proponents of PPR is supposed to conclude that the IR view cannot adjudicate among preferences and in fact instrumental reasoning only *starts from* the point when an agent has decided which end is most important.[33]

Of course there is a trivial way in which agents prioritise their ends and that is when they are in situations where it is possible to satisfy one end and there is no possibility of satisfying others. For example if Jane wants to travel the world for ten years, find a cure for cancer, and become an astronaut then we can probably safely assume that these three ends might be incompatible. The prime years for travelling with a backpack on a budget and for the gruelling astronaut training are roughly the same and not many people go into medical research in middle age. On an IR view, as proponents of PPR would have us believe, Jane has no way of choosing among those ends and any method she does use will be non-instrumental. But it is trivially true that Jane might choose one of the three just because one chance emerges rather than the others and so we can say that it is true on an IR view that proximity to a desired end (for example, when an agent has the available means to satisfy one particular end *right now*) can bring about the satisfaction of that end. In cases such as this we can say that

---

[32] Of course not all our ends are unmotivated. The ends we have because they are the means to some further unmotivated end are *motivated* ends.

[33] Brandt (1979) and Gauthier (1986) are accused of this by Korsgaard (1996).

the agent chooses to satisfy the end but does *not* thereby resolve the conflict among her ends.

Because of problems about conflicts of ends IR theorists tend to claim that an agent is rational if she takes heed of some coherence conditions and applies them to her set of ends or preferences. The axioms of expected utility theory come to mind here. However IR theorists stand accused here of importing and applying normative principles to enable them define an agent's set of ends. The ends themselves may still be unmotivated but in order to form a coherent set they must undergo a process the end result of which will leave the agent with an acceptably coherent set of ends. Thus the IR view appears to require the agent to put her preferences through a process that is itself a normative procedure. Any agent who is to be considered (instrumentally) rational will have a set of preferences that is coherent and this set will be part of a larger set that includes her total preferences.

Of course admitting coherence constraints on preferences on a theory of instrumental reason is moving beyond the position Hume originally outlined, but proponents of the IR view see the necessity for such constraints and have accepted that there is more to the IR view as a theory of practical reason than merely working out the best means to given ends.[34] However PPR proponents deny that such a move beyond Hume will lend any greater plausibility to the IR view. Let us take the constraints of the expected utility view as an example. To decide which of two actions is more rational to perform we have to decide to what degree we allow proximity of the satisfaction of one preference to be relevant to the choice of action. For example, I want to go home to check my tomato plants and this action will give me an expected benefit in six weeks time. I also want to stay here and write and this action will give me an expected benefit tomorrow when I will be relieved that I have not fallen behind

[34] For example, Nozick has admitted as much in (1993) p.40.

in my daily writing quota. Which of these actions is more rational? To what extent are we supposed to allow considerations about future satisfaction of preferences override more immediate concerns?

The most common or commonsense answer is that we should be biased in favour of our future, long-term preferences. But, PPR theorists ask, is it possible for an IR theorist to justify such a claim? A possible IR answer is that this bias can be explained by our having a second-order preference for when we would like our first-order preference to be satisfied. By claiming that it is no more than a second-order bias we avoid having to admit that it is rationally required and thereby we can deny it normative status. However PPR supporters attempt to deny that the IR view really can support such a response. As Hampton points out, for example, it is true that we seem to regard the person who only thinks of their immediate ends as somehow wrong or irrational and it does ring true that most IR supporters see the focus on one's long-term goals as prudent and praise-worthy.[35] But, Hampton argues, the objection to the IR view taking on this bias in favour of long-term goals is that it just cannot be made consistent with a view of practical reasoning as purely instrumental and so if the defenders of IR want to be consistent then they must accept that the agent who favours her short-term goals is as practically rational as the agent who favours her long-term interests. Of course the PPR defenders are throwing down a challenge here. IR supporters will either have to admit that they do hold certain normative commitments that can explain their bias in favour of long-term preference satisfaction or they have to accept that the agent who rushes off to check the tomato plants instead of working towards her long-term goal is just as rational as the agent who diligently works to achieve a long-term goal.

---

[35] Hampton (1998) p.180-1.

Even if IR defenders did bite the bullet about the ranking of ends and accept that no agent can be called irrational because agents are always rational relative to their own ends and beliefs, a substantial problem remains with accounting for preference change over time. This problem centres around the issue of whether an agent should act on her *current* preferences or her *future* ones. The challenge here is to give an account, consistent with IR, explaining why we ought to include or exclude future preferences from our deliberations about how to act. Take an agent who has a current preference to get up early in the morning. She sets her alarm for early but the problem is that when she is in bed in the morning she will have to consider what her *then* current preferences are in order to decide how she should act. The trouble is that the agent might decide that her current preference is to stay in bed. But if agents have to consider their current *and* future expected preferences then it is clear, so the objection goes, that some criteria for evaluating preferences is required. However none can be found unless it is recognised that whatever criteria are used are *normative* in origin. If the IR view is supposed to have the resources within it to distinguish between different types of practical reasoning – be it short-sighted, prudent, reckless etc. – then proponents of this view must recognise that none of these types of reasoners can be picked out in these evaluative ways unless the view presumes prior normative commitments. And if there are such commitments, we are told by the critics of IR, then it is only a short step from the recognition of this fact to an inevitable further recognition of the supposedly deep inconsistencies to be found in the IR view.

What the IR view lacks then, according to its critics, is the ability to give us a reason to pursue our ends in the first place. This means that the instrumental norm cannot be the source of normativity for us. It can allow us find explanatory reasons for our actions but it can never be a source of justificatory reasons because it has no jurisdiction over our ends. If it is true that the IR view cannot account for all the sorts

of actions we typically come across then we are going to need to find a theory of practical reason that can explain all the phenomena. The PPR view sees itself as precisely this alternative theory; but the question we have to ask ourselves is whether proponents of this view can convince us that the only way to have any justificatory reasons for our actions and to be able to evaluate our ends is to adopt a view of practical reason that is derived either from a pure principle of practical reason or from a conception of our own agency as free, autonomous agency, in the sense Kantians like to spell out.

What the PPR defenders must show us is that there are a number of crucial errors being made by the IR view. We have been discussing the first such apparent error – the claim that IR defenders have not distinguished the instrumental principle from the principle of prudence. Korsgaard claims that the instrumentalists have blurred what should be a clear distinction by taking it for granted that a person's overall good is what they *really* want.[36] This allegedly unjustifiable conflation is supposed to be used by the instrumentalist to avoid the issue of the source of the normative status of an agent's justificatory reasons for her action. If Mark decides that he really wants to spend all this month's earnings on one glorious day out then he is doing what is in his best interests and therefore all he has to do is put some quality time into thinking about how to really spend his money well. The difficulty for Mark is that he cannot evaluate his ends, he does not have the capacity to do anything else other than believe that his end is good *because* it is chosen by him. Since IR defenders accept that agents can be relied on to be motivated to take the means to their ends the PPR view claims that the instrumentalists have helped themselves to the assumption that the instrumental principle does not require any further authority (i.e. *normative* authority) for its justification. The practically rational agent on the IR view is one who is caused to have

[36] Korsgaard (1997), p.218.

motive to act when she approves of certain reasons for action that are in accord with her existing predispositions. But there is no *necessity* for the agent to act.

On a traditional reading of the IR view necessity simply refers to *causal* necessity,[37] which means that we can infer that a certain effect will follow from a particular cause. When it comes to practical reasoning then all we can say is that an agent is in fact caused to act by the recognition that an action will promote her chosen end. But there is no sense in which the agent takes any specific guidance from practical reason since this amounts to the claim that we can reliably say that if we know what an agent's ends are then we can predict that certain action will follow. So the objection here is that there does not appear to be any such thing as practical reason available to give any sort of guidance to agents. Instead they have to rely upon their desires and predispositions in general to provide reasons to do one action rather than another.

It is conceivable however that for all that has been said so far the instrumental principle and guidance by pre-existing desires might form the basis for an acceptable theory of action. But the objections to such an IR position run deeper than this. As already mentioned another central objection to the IR view is that the instrumental principle cannot give an agent reason to do any specific action because it only tells us to take the means to our given or chosen ends. If it is to form the basis for a theory of any sort of practical reasoning (in the broadest possible sense of the term) then there

---

[37] Hume *Treatise* (I.iii.6). However I do not think that the traditional criticism of this view fully takes into consideration Hume's view on *reliability*, introduced at T404. Also, when it comes to practical reason, Hume argues that *a priori* judgements cannot affect action in any essential way; but this does not mean that he, or someone agreeing with this, must say that sense is the sole origin of moral conduct. In the latter sections of Bk III Hume does not exclude reason from influencing moral conduct (and this is even more pronounced in the *Enquiry*). Importantly, in his definition of approval (the ground of moral obligation) and disapproval Hume says that it is not just a matter of sense (although it can be direct too, if we have a 'proper relish' of the object) but that approval also results from reflection and judgement. Approval is not the same as 'liking' because it arises from an impartial consideration of its object. (T180) Hume wants it to be possible that we can figure out whether certain value judgements are true by measuring them against a standard which is common to all people of the same susceptibility. For example, see T 277-8 for comments on 'steady general points of view.' See Rachel Kydd (1946); Annette Baier (1991) for careful readings of Hume, an adequate interpretive exploration of which is outside the scope of this thesis.

must be some view about what our ends in fact are or about how we figure out what they are. The objection is that for the IR view to be coherent it desperately needs one assumption that it just has no entitlement to: that our ends are set by, or in some way identical to, what we desire.[38]

If the instrumental principle is the only principle of practical reason then no agent can say that she has a *reason* to pursue her ends. Rather the objection is that all the agent in question is entitled to say is that she is *going* to pursue her ends because she is motivated by her predispositions to do so. The IR view would only be entitled to claim that an agent has a *reason* to act if it included some further principles designed to permit the evaluation and selection of ends. If there is a reason to pursue an end then there will be at least a prima facie reason to pursue the means to that end. This however is the formulation of the instrumental principle most objected to.[39] The apparent incoherence in the IR view is this: if reason is denied a role in determining or evaluating ends then we are not entitled to claim that an agent who is about to pursue an end thereby has a *reason* to take the means to that end. The most that can be claimed, according to the objectors, is that the agent with a certain end is (in all likelihood given the reliability of the relation between cause and effect) going to or will pursue the means to that end.

Now a reason is something normative and the argument against the IR view claims that we are not entitled to derive something normative (the *reason* to pursue the means) from something non-normative (the fact that the agent has certain ends). But since IR theorists are supposed to deny reason any role in the evaluation or formation of ends their objectors claim that they seem to attempt to derive reasons from facts

[38] Of course the IR defender can have a theory about how we come to have certain desires. We can, perhaps, form desires on the basis of social approval of particular activities, so for example, Mary comes from a society that approves of self-improving activities and disapproves of laziness and inactivity in general and consequently Mary's pre-existing desires are formed by these social biases. The point is that the IR theorist is not forced to claim that pre-existing desires are unmotivated by any outside influences.
[39] See for example Korsgaard (1998), p.223, McDowell (1978), Hampton (1998), ch 5.

about what agents *are about to do* or from what we can predict with some certainty that they will do. In this case the instrumental principle could be read in a *justificatory* sense: (1) if you have an end then you have a reason to pursue the means to that end. However the argument we are examining suggests that the instrumental principle can only be read in the following *descriptive* sense: (2) if an agent has an end then we can predict with some certainty that she will pursue the means to that end.

To briefly sum up the argument against IR thus far, the accusation is that the IR view can only argue for (2) consistently. But if IR can only argue for (2) then those who defend IR have the following problem. In order for IR to claim an agent to be weak-willed the agent must fail to take the means to her chosen end (by knowingly acting on an already overruled reason). But to say that this agent is weak-willed is to assume that the instrumental principle is binding in the sense of being normatively necessary. But according to the objection IR cannot have an account of the instrumental principle as normative. Therefore the descriptive reading (2) of the instrumental principle does not permit an explanation of weakness of will (that is, it cannot be made consistent with the internalism requirement). If for the reasons we have explored neither the ranking of ends nor preference change over time can be adequately explained (other than in terms of the sort of successive causal interplay of different states of the agent that Hume describes at the end of section II.iii.3 of the *Treatise*) then it is not possible to distinguish (in terms of their rationality) between agents who act on their immediate ends and those who act on their long-term ones. So by implication the weak-willed agent cannot be clearly distinguished from the 'rational' agent. IR cannot account for weakness of will because it cannot distinguish between what an agent's ends are and what she actually pursues. Therefore IR cannot meet the requirements of a

theory of practical reason.[40]

### 3.3 The Necessity of Rational Requirements

As we are beginning to see the deeper problem involved here is that the IR view is in danger of not being able to hold any position at all about what practically rational action is. If the instrumental principle can give us no authoritative guidance about how to act and if there is no way in which we can evaluate our ends then we can indeed have some sympathy with those who find this a nihilistic and dangerous prospect.

The objections we have been examining so far centre on the failure of the instrumental principle to act as a normative principle that could guide practical reason. At this point however, an IR defender might want to suggest that it is not the instrumental principle only that guides practical reason, but the principle of *prudence* also plays this role (that is, that it is rational to act in a way that safeguards or promotes one's future or long-term self-interest). PPR defenders have also examined this possibility and not surprisingly they argue that on an IR view prudence cannot be considered a rationally required *principle* and so cannot be considered normative. Since neither prudence nor instrumental reasoning can be rationally required on an IR view then irrational action cannot be explained on such a view. Further, if it cannot explain this then, so the argument goes, it will not be possible to understand how an agent can be practically rational in any sense. In other words, the failure to have any rational requirements at all upon action results in a theory which effectively reduces talk about actions to talk about the motivational efficacy of desires and predispositions

[40] This threatens rational agency for if the agent is not practically rational then the fear is that she cannot be self-governing and so she will be less like an agent and more like a bundle of desires or dispositions (that is, a wanton).

in general. The danger is that agency itself is threatened on such a view. If the agent is not a reflective endorser of her own reasons for action, if she is not practically rational, then the agent will not be autonomous and therefore will be incapable of self-governance. There is a lot at stake here.

Korsgaard presents an argument the conclusion of which is that Hume cannot account for irrational action unless he re-characterises not only the instrumental principle but also *prudence* as rational requirements on action.[41] She argues further that once it is recognised that these *are* indeed rational requirements, a satisfactory explanation of practical reason will only be reached when we understand the extent to which practical reason is nothing other than autonomy or self-governance. In other words, we start off with Hume and we end up with Kant. The key to this argument is the treatment of practical agency and here the problem of weakness of will is central. The most important step in this argument involves showing Hume's (or the IR view's) failure or inability to differentiate rational from irrational action. The only solution offered is to accept the inevitability of the alternative neo-Kantian "ethics of autonomy."[42]

### 3.4 The hypothetical imperatives: prudence and the instrumental principle

Returning to Hume's original argument for the IR view in relation to his well-known conception of reason 'as a slave of the passions', we get a fairly unambiguous claim about the hypothetical status of prudence. We saw this famous statement about preferring one's own acknowledged lesser good and he goes on to clarify the point:

> 'Tis not contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter. A

[41] See Korsgaard (1986) and (1987).
[42] Korsgaard (1996), ch.1.

> trivial good may, from certain circumstances, produce a desire superior to what arises from the greatest and most valuable enjoyment; nor is there anything more extraordinary in this, than in mechanics to see one pound weight raised up a hundred by the advantage of its situation.[43]

But we do in fact tend to act prudently according to Hume. We have an 'original instinct' or psychologically basic constitutive element that is in us by nature, that explains our tendency to perform prudent actions.[44] Such actions manifest themselves as a tendency to desire what will be in one's overall best interests combined with an aversion to things that will prevent us from achieving those interests. This tendency is described by Hume as a 'calm passion' – something we know more from its effects than its felt nature. In the well-known section 'Of the Influencing Motives of the Will',[45] Hume tells us that we have a tendency to confuse the operation of the calm passions with the operations of that other calm faculty, theoretical reason. It is because of this confusion that we think that prudent action is a special sort of *rational* action. However Hume is adamant that when we are not acting under our 'general appetite to the good'[46] we might act against our overall interests, yet even if we do, we cannot be accused of acting irrationally since acting according to the calm passion is not 'rationally required' in the first place.[47] Now acting to further one's best interests involves fairly clear beliefs about what those interests are and how future events might effect the furtherance of those interests. A belief for Hume is forceful, clear and vivacious in a way that is very similar to an impression. For example, if I believe that the drink in the cup is petrol rather than water I will avoid it in much the same way I would if I tasted it and recoiled in horror. In this way, according to Hume, nature has

---

[43] *T* II.iii.3, p.267.
[44] *T* 438.
[45] *T* II.iii.3, pp.265-8
[46] *T* II.iii.3, p.268.
[47] Except perhaps only indirectly, on moral-pragmatic grounds having to do with the desirability (on Hume's view) of having certain stable, pleasing, and productive character traits, which the predominance of the calm over the 'violent' passions would facilitate. This is a sense of 'reasonableness' which is not the sort of ground of obligation that the PPR view is demanding, however.

constructed us to be motivated by our beliefs about the future consequences of our actions.

So prudent foresight is motivational in itself, we might say, and does not require an *extra* desire beyond our 'general appetite to the good and aversion to evil' to motivate us to act on it. But the prudent agent who acts on her belief about her future desires or interests is not rationally required to so act, nor could she be criticised as irrational if she failed to so act; for all we can say in such a situation is that she does not have the right beliefs about her future desires or that she does not have an operative calm passion predisposing her to act on her desire for her own good. In other words, all we can say about the agent in this position is that she either makes certain mistakes or has an undesirable structure of passion-based predispositions. But we cannot say that she is *irrational*. If it is just a matter of our natural character to have the capacity to be motivated by our beliefs about future matters then we cannot say that this amounts to a rational requirement of practical reason. It is important to remember here that Hume believes that while a belief can be as lively as an impression, any agent who has a belief about her future desires or what will be in her interests next year for example, must have a present general desire for her own good also. Without this no prudent action will take place.[48]

Hume certainly does believe that a more immediate desire can overcome one's general appetite for the good, yet it is this stage of his argument that seems to cause the most trouble. Korsgaard, for example, interprets this as the central claim in Hume's argument that prudence is not a rational requirement of practical reason. She suggests that on Hume's view if prudence "...fails to motivate us, the principle of prudence fails to meet the internalism requirement, and so cannot count as a rational principle."[49]

---

[48] This point is somewhat passed over in the rush to interpret Hume as having at least one type of belief that is motivationally efficacious in itself. For example see Korsgaard (1997), p.226.
[49] Korsgaard (1997), p.226.

Korsgaard understands the internalism requirement as a necessary requirement on practical reasons that motivate the agent to act. In order to fully explain an agent's actions, reasons must be motives (but the reverse does *not* hold unless we say that motives can be *prima facie* reasons but not *all-things-considered* ones). If this was not the case then the agent could ask why she should do what she knows she has reason to do.[50] Korsgaard argues that Hume believes that if prudence occasionally fails to motivate us then the principle of prudence does not meet the internalism requirement and if it does not meet this particular requirement then it cannot count as a principle of practical reason.

According to Korsgaard this position is based on a misunderstanding of the internalism requirement, because she will only allow this requirement to specify that practical reasons must motivate us only in so far as we are rational or "susceptible to the influence of reason."[51] Her view on the internalism requirement is clear. Because the internalism requirement does not require that agents *necessarily* be motivated by their reasons, any kind of Humean scepticism about the capacity of practical reason to motivate an agent to act is misguided. We should only be sceptical about a kind of practical reason that is not defeasible in the way she believes it is. A more plausible and robust internalism about reasons *is* defeasible and of course it is necessary to explain what is involved in acting on reasons (both from the first and third person perspectives). Korsgaard's view in this connection is that Hume misunderstands the internalism requirement by believing that for some consideration to count as a practical reason it must actually succeed in motivating the agent holding it to act.[52] Since we

---

[50] This is more specific than our general characterisation of the internalism requirement as discussed in the Introduction.

[51] Op. cit.

[52] Some readers may have doubts about whether Korsgaard is getting Hume right here, since he clearly talks about what would move an agent "in a cool hour" or "from a general point of view." Annette Baier and Barbara Herman pressed Korsgaard on this point and in reply she tries to distinguish between Hume's view and the views of what she calls contemporary 'empiricists' who claim a Humean heritage.

hold plenty of considerations that never do succeed in motivating us these cannot count as practical reason; in other words, if any considerations are to be *rational* requirements they must motivate the holders of them, but since they do not always or necessarily do so then they cannot be rationally required.

There is one supposedly clear implication of this (IR) line of argument that Korsgaard assumes we will accept, and that is that the agent who fails to act on a certain consideration can never be accused of acting irrationally. This is not just because practical rationality on the IR view only generates instrumental reasons to act. What we are presented with here is the claim that for Hume or any IR theorist, their view of practical reason makes it impossible for any agent to act irrationally or demonstrate weakness of will. If Jane was not motivated to go to the slimclub meeting even though she recognised this as a consideration worthy of action, then on Korsgaard's reading of Hume we have to say that Jane *must not really have had a reason to go in the first place.*[53]

What is in this way being attributed to Hume and the IR view is an understanding of the internalism requirement that is not compatible with its being defeasible by any of the usual human failings. Now any theory of action needs to have the conceptual tools to explain the mistake an agent is making who fails to choose to perform an action that will lead to her greater good. If the IR view can only tell us that this agent must not really have or have had the motive to seek her greater good (i.e. she has no motivation of prudence at all) then it would be difficult to disagree with Korsgaard regarding the need to discard such a position. There does not seem to be any plausibility to a view that can never allow an agent to do anything wrong or irrational. Korsgaard's claim, in a nutshell, is that the IR argument is circular. It assumes that

---

Korsgaard argues that for Hume the desire to take the means to our ends is a calm passion, not a normative principle. Anyone who takes it to be such a principle is making a mistake.

[53] Take it as read that Jane was able to go, that she didn't have a more important appointment to honour etc.

reason cannot motivate in order to show that because sometimes reasons do not actually motivate agents to act therefore reasons alone are not motives. In other words, scepticism about the motivational power of reasons depends upon the prior assumption that reason alone cannot motivate.

The IR theorists have tried to show that scepticism about the motivational power of reasons actually serves as proof that reason alone cannot motivate. Informally, the IR view is accused of looking something like the following:

(1) If it is true that considerations in favour of any given action do not always motivate agents to act then it must be the case that considerations for action are not necessarily motives for action.

(2) Considerations do not always motivate agents to act.

(IntC) Therefore, reasons for action are not necessarily motives.

(4) If considerations in favour of actions are not motives for action then motives for action must exist independently of practical reason.

(5) If motives exist independently of practical reason then for any action an agent will need a consideration in favour of the action and a motive to act on that consideration.

(C) In order to perform an action an agent requires a consideration in favour and a separate motive to act according to her consideration for action.

Now line (2) is where the supposed error lies since this is, we are being asked to believe, a rather flagrant begging of the question. On this construal of the IR view the conclusion to be proved is assumed at (2) and so what ought to be the conclusion – the claim that considerations in favour of an action are insufficient in themselves to motivate an agent to act – is used to reach the conclusion that pure practical reason is impossible.

There are a number of accusations being made here. In the first place the IR position is thought to rely on a conclusion it cannot reach without begging the question, as outlined above. Secondly, the problem of weakness of will is considered illegitimately disposed of with a denial of the possibility rather than an explanation of the phenomenon. Korsgaard and Hampton argue that this latter problem is a direct consequence of the former error. However this is a matter we can dispute. It may or may not be the case that Hume himself committed these offences but for our purposes what we have to examine is not the possibilities open to Hume interpretation, but whether our contemporary IR theorist can be accused of the same errors. What I want to suggest in outline here is that these errors are not as inevitable as Korsgaard would have us believe as they just are not the unavoidable conceptual pitfalls of the territory she claims them to be.

As we have seen the second major stumbling block appears to be the analysis of prudence. This objection to the IR view is that in order to say that an agent performed an act that at the very least *she herself* could consider imprudent, prudence must be a rational requirement on practical reason. For example, if the agent we met earlier, John, knows the most prudent course of action for him is to try to launch a career in the civil service then any action he engages in that thwarts this is imprudent and as such we can say that it is practically irrational (self-defeating or weak-willed) of him to go off and try heroin instead of sitting the exam. This (imprudence charge) then is supposed to be the only way an IR view can criticise John for his activities. The alternative certainly seems unappealing. But if we claim that prudence is not a rational requirement on practical reason we have to say that because John tries the heroin, that must be what he really wants to do. If this is what he is *going to do* then we have to assume that he has the relevant and operative dispositions pushing him to perform this action. If this is what John is really predisposed to do then how (on the IR view) can

we accuse him of being irrational? There are just no conceptual tools available on this picture of practical 'reasoning' to criticise John in any way he ought to find *rationally* binding. John is instrumentally rational because he takes the means to the end he is going to pursue as opposed to the end he thinks might be in his best interests. We may hypothesise about why John does this and there are many things we could say, such as for example, that he is short-sighted, that he will repent at leisure, that he is foolish, that his 'violent' passions tend to win out over his 'calm' passions, etc., but we cannot say that he is irrational or weak-willed. Any agent can be mistaken about what the means to a given end might be and so in that case they could be called *instrumentally* irrational; however this agent still has the opportunity to take the correct means once her mistake has been pointed out. In this case, then, the IR view does not accommodate the undeniable case of the 'clear-eyed' practically irrational agent[54] who is knowingly acting against her own interests, hence such an agent could not actually be called irrational.

What worries the opponents of the IR view is that it implies that no agent can be called irrational: for it is not possible for anyone to violate the instrumental principle because all agents will of course take the means to the ends they are actually about to pursue. If the instrumental principle was considered a rational requirement (or even the only rational requirement) on action then the critics might be a little less troubled, for at least we could say that any number of a broad collection of dispositions was interfering with the proper process of practical reasoning and action. John above might be afraid of settling down, he could be lazy or perhaps dazzled by illicit excitement, and any of these could explain why he violates the instrumental principle. But if the IR view does not permit this principle to act as a rational requirement then there can *never* be any

[54] Much of the literature on weakness of will from the mid-twentieth century focuses on the distinction between 'clear-eyed' akrasia and other forms of irrationality. The clear-eyed akratic is one who we take to be the standard weak-willed agent, that is, one who *knows* that what she is doing is contrary to her own acknowledged all-things-considered reason for action.

sense in which John's action is not instrumentally rational. The problem with this is that the instrumental principle becomes no more than a claim about the 'constant conjunction' of means and ends and this will be simply the result of the observation of behaviour from a third-person perspective. In other words, anything an agent does is the means to the end she *is* pursuing and this seems to be just an observable fact about agents. But if taking the means to the end is just something agents predictably do, then they cannot be guided by the instrumental principle. It cannot, therefore, be *normative* for any agent. As we shall see later, however, to deny that something can be a rational requirement on practical reason does not necessarily mean that it cannot be normative. There appears to be a rush to infer the latter from the former, but what needs to be more closely questioned is the assumption that the source of normativity is to be found in the recognition that there are *rational* requirements upon one's reasons for action.[55]

It is worth pointing out at this stage, however, that the opponents of the IR view are correct in insisting that the instrumental principle has no role to play in determining our ends. This just must be true, for how can a principle which says *only* that we take the means to our chosen ends implicitly say anything about what those ends are to be? However I do think we can plausibly disagree with the rest of this objection. Hume is accused[56] of identifying an agent's ends with what she wants most and the only way to find out what she wants most is to see what she *actually does*. If someone *always* (and without exception) pursues her chosen end then she will never violate the instrumental principle, for as we have seen there is no way she can do anything but take the means to the end she is, or is about to, pursue. But Hume does not claim (as advocates of the 'rational choice' version of the IR view do[57]) that it is rational for everyone to seek their overall good, for if he did then he would be claiming that agents feel obligated to

[55] We will be dealing with this issue more particularly in the next chapter.
[56] Korsgaard (1997), p.230.
[57] For example, see John Elster (1986).

drop the pursuit of ends that are inconsistent with what they believe to be their overall good, and this would all too obviously be importing the rationality of prudence back where it supposedly does not belong.

According to Hume we *want* those things that constitute our ends and not just those that are consistent with our overall good. But if we want all those things then we will, of course, always take the means to our ends and so we are always instrumentally rational. What is probably evident at this stage is that on Korsgaard's picture of a standard version of the IR view the instrumental principle looks much less like a *principle* and more like a predisposition of agents. To claim that it is a principle is to say that it *guides* action or more precisely that it functions as a standard an agent can appeal to when deciding how she ought to act. However if taking the means to one's ends is just something agents *do* and cannot but do then the instrumental reasoning is misdescribed as an action-guiding principle and, indeed, as *reasoning*. Rather what we have is action brought about by the predisposition of agents to take the means to their ends. Further, we would have to say that this is a commonly found or widespread disposition with no more necessity to it than that. If an agent fails to act to achieve her desired end, then she is just making a mistake – nothing hangs on whether she actually pursues the prudent course or not. If she does, it is because she prefers it and if she does not it is because she yields to other passions. But she still cannot be criticised for failing to take the means to an end she simply prefers.

Once again, as with the instrumental principle, prudence is not guiding her actions but is only, as it were, a general term for that factual truth concerning agents' predictability which the instrumental 'principle' describes. So it appears that the agent cannot do anything other than act the way she acts. What causes concern among opponents of the IR view, as we have seen, is that the IR representation of practical agency seems to paint a picture of the typical agent as one who never violates any of

the principles of practical reason because it is impossible for her to do so. If the agent simply acts in accordance with and oblivious to these predictive or explanatory principles then it must be the case that she is in no way *guided* by them; and if she cannot be guided by such principles then we are told that they just cannot be normative for her.

Behind the fear of this implication, however, is a more substantive disagreement concerning the nature of agency itself. If there is no distinction between an agent's ends and the act she eventually performs then one might be tempted to object that the agent seems to be no more than a bundle of desires, beliefs and other dispositions. For example, we might say of John that the IR view only sees him as the locus for the operation of certain beliefs and desires, such as the desire to try heroin or the belief that the civil service makes for a good career. But where is John *himself*, asks the objector. As Korsgaard argues, for example:

> [d]esire, fear, indolence, and whim shape the Humean agent's ends, and, through them, her actions. When her passions change, her ends change, and when her ends change, so do her actions. We can explain everything that she does without any reference to *her* at all.[58]

The *agent* is not identical with the sum of its desires, beliefs and general dispositions to act on this view. Rather the agent is supposed to be that element that does the reflecting and choosing while being guided by normative principles.

Now such a bold statement of an anti-rational conception of instrumental reasoning strikes fear into the hearts of many because of the loss of normativity supposedly entailed here. If there is no normativity to be found at this most fundamental level of action then the fear is that no action an agent performs will be open to criticism. If there is no way in which what I do is ever even instrumentally irrational (or not fully rational) then it is difficult to see how any less obviously binding

[58] Korsgaard (1997), p.233.

principle such as prudence could be in any way normatively binding. If simple means-ends reasoning is in this sense merely dispositional (that is, a 'calm passion' we have by original constitution[59] why should we assume that other considerations provide normatively binding rational requirements on action?

In sum, then, the real fear is that the IR view ultimately leads to an inevitable rejection of normativity altogether. If there is nothing normative about practical reason then there is no such thing as practical reason at all, for such reasoning is supposed to provide the agent with a network of norms by which she can decide what she *ought* to do. If all we can do is maintain a degree of predictive success about how agents will act given their dispositional character then we will have succeeded in explaining away entirely the evaluative dimension of our lives, for we will not be able to distinguish between what I will, or am about to do and what I think is best to do.

### 3.4 The IR View and Irrational Action

Now as we have already seen, any view which attempts to deny irrational action[60] runs into fairly insurmountable obstacles and if nothing else, is deeply counter-intuitive. So to this extent we can agree with those objecting to the IR view on this basis. If the IR view does indeed entirely fail to allow or account for irrational action then it is a view we should reject. However if we can show that irrational action *is* in fact permitted and does not have to be denied then we shall have much less reason to reject the IR view in favour of the PPR view. What IR theorists are supposed to find impossible to achieve is any sense of practical reason providing general principles or imperatives which can

[59] Korsgaard, (1997), p.229.
[60] Even action that is irrational only in the weak sense that it is contrary to how the agent reports her own judgement about how she should act.

*guide* an agent to act in one way rather than another. If the IR view is purely descriptive of how agents *do* act then it is not possible for them to be guided by practical reason and if they cannot be guided then they cannot *fail to act* according to the dictates of practical reason.

The IR theorist is thus accused in this more general way of not being able to explain how reasons (even just instrumental or prudential ones) have *authority* for the agent. If reasons just are identical with the considerations we hold in favour of actions then it appears that reasons have no status other than as the agent's desire or predisposition in favour of X. For example, if we say that John is motivated to go to the cinema in accordance with his *reasons* for doing so, then when asked what these reasons are all that can be said is that John wants to go to the cinema or that he somehow feels that this is a worthwhile experience etc. The point is, if we accept this criticism, John's reasons are not considerations he recognises as normative for him. They cannot be, because all John is telling us, as it were, is how he is motivated by instrumental thoughts; he is not able to explain how such thoughts could be guides to, or rational requirements on, his action.

The IR theorist is therefore supposed to be missing the crucial point that the descriptive can never amount to the normative; or put the other way, that the normative cannot be successfully reduced to the descriptive.[61] Fuelling this claim is the view that practical reasons are attempts by agents to justify their proposed actions. To look for a reason to act is to search for a *justification* of that action, whereas facts about ourselves – cultural, psychological, biological etc. – can never satisfy the stringent requirements necessary for considerations to function as justifications. To search for a reason for action is to search for a normative answer to a normative question. Predispositions may actually guide our actions in some situations but since it is always possible, we are told,

[61] See also Nagel (1997), Ch.6.

to normatively reflect on our dispositions to action, giving an account of my action in terms of my desires, for example, can never justify my action as something I did as a result of my thinking that I *ought to do it*. For this reason also the IR theorist is not supposed to be entitled to have a role for the instrumental or prudential *principles* in the explanation of action. There is no alternative on the IR view except to see willing or choosing the means to a chosen end as descriptive of performed action or action currently being performed. It has to read something like this: choosing, desiring or willing the means to an end is the same as actually pursuing the means to an end.

Korsgaard's alternative PPR view sees the instrumental principle as *constitutive* of any act of volition or of willing an end. In other words, if something is to be considered a *reason* for action then it must already conform to the instrumental principle or it is not a reason for action at all – willing the end just is to will the means. Consider the following example. Mary chooses to throw paint at women wearing fur coats. Her goal or end is to free animals from persecution by fur-traders. Now Korsgaard's view implies that the IR view must claim that someone like Mary has her reason for action and *then* she must consider whether that reason conforms to the instrumental principle.[62] So Mary's practical reasoning goes something like this: "Throw paint at women in fur coats. Is throwing paint at women in fur coats a means to my end of freeing animals from persecution?" How she answers this question will determine her action because the instrumental principle is acting as a restriction to be *applied to* a reason she already has. However even if Mary goes ahead and throws the paint all we can say is that she must therefore have *really wanted* to do so. If we take her action as an indicator of her strongest desire or of what she had most reason to do then it will be impossible for Mary to violate the instrumental principle, and this being the case it cannot be a requirement of, or guide to, action. The instrumental principle

[62] Korsgaard (1997), p.244.

then, cannot play a normative role in practical reason. Or at least this is how the argument goes. But I think we can save the IR view from objections of this nature, and we will canvass the possibilities later. For the moment however, let us continue our look at the objections.

If conforming to the instrumental principle is constitutive of having a reason to act then to have a reason will just be to have a reason to bring about one's end and so one already has a reason to take the means to that end. If Mary has a reason to throw paint then she knows that she cannot go further and ask whether doing so is a means to her end. To say that the instrumental principle is constitutive is not supposed to be a claim about necessity but about *normative necessity.* The most familiar way to explain normative necessity is in terms of personal commitment.[63] Choosing or willing an end is nothing more than committing oneself to bringing about that end. Having a reason then is, on this view, an entirely first-personal act and rules out altogether what is taken to be the IR view of reasons as third-person descriptions of the pursuit of ends. Korsgaard conceives of this personal commitment as the giving of a law or principle to oneself and to act in accordance with this law is to be engaged in governing oneself.

So our agent, Mary, if she has the protection of animals as her end will, if she is a self-governing agent, commits herself to taking the means necessary to realise her goal. This is supposed to eliminate the series of questions agents could ask themselves about their actions and reasons for doing so, in the following way. On the current interpretation of the IR view, Mary knows she wants to save animals so she thinks she should throw paint at those who are responsible for keeping the fur trade alive. But since knowing her end does not (contra Korsgaard) commit her to taking the means she can ask herself whether her action really is a means to her end and even if she thinks it

[63] Korsgaard (1997), p.245. In chapter five we will be arguing in favour of an alternative account of commitment – one that does not assume that one's all-things-considered reason for action is necessarily motivating.

is she can still wonder whether she should do it. She is not, after all, committed on pain of irrationality to take the means to any given end. If she decides not to throw the paint then we are entitled to conclude only that she must not *really* care so much about animal welfare.[64] And so we are back to the same problem – the IR view cannot, so the objections go, account for irrationality, it can only give a descriptive account of reasons and is not entitled to claim that the instrumental principle has any normative authority. Claiming that someone is going to or is about to pursue an end does not make it normative for that agent. If we say that Mary is *going to* throw paint this does not give her a *reason* for her action because all it says it that Mary is pursuing a particular end. This IR formulation supposedly fails to capture the sense in which Mary's choosing or willing her end is a normative commitment on her part. If she wills to save animals then she also and at the same time wills the means to that end because what she has committed herself to is making it a binding principle *for her* that having an end involves *obligating oneself* to taking the means to that end. On a common interpretation of the IR view[65] obligating oneself amounts to no more than a contingent predisposition to do what leads to our ends. The most we can get from any of this is contingent authority – it can be true that an agent *ought* to do A only if she happens to have the relevant disposition to do A.[66]

So much for a common and important objection to the IR view and we will return to these issues later. But perhaps the most serious objection thrown at the IR view concerns something we have only discussed briefly thus far, and that is *agency*

---

[64] We are assuming here that she is not just making a strategic move, that is, she might decide not to throw paint because her intended victim is also walking a dog that might be traumatised by the event, or she does not want to risk arrest because she is getting married tomorrow. We are assuming then that she does not decide that she has another, more immediately important end to realise.

[65] For example, Hampton (1998).

[66] Even if we accepted these objections it might be suggested that there still might be a way in which the IR theorist can talk of instrumental *reasons*. If we argue that talk of instrumental reasons is a useful framework which can be cashed out ultimately in terms of some sort of error theory then it could be possible to maintain some sense of practical *reason*. However this does seem too high a price to pay since it eliminates the reality of the normative – precisely that which contemporary IR theorists acknowledge has a crucial role to play.

itself. If there is no practical reason in the sense that there is nothing it is like to act rationally then, as this objection would have it, a question arises concerning the identity of the agent. If we take Plato's charioteer as a metaphor, how, on an IR view, can there be a charioteer at all? If there is no practical reason how can there be any practical *agency*? The worry is that if the agent just *happens* to have the ends she has and just *happens* to have certain predispositions then it becomes difficult to see how the agent could be anything more than simply a combination of desires or forces pushing and pulling without any central control. The picture would better resemble the chariot and horses operating as something more like a complex automated system than an agent who can choose whether to act on her ends and detach herself from and reflect upon her desires and impulses.

The minimum condition for agency on this view, is conformity to the instrumental principle. To be more precise, this conformity is *constitutive* of having a will or being capable of acting for reasons at all. Without this conformity it is not possible to distinguish between the practically reasoning agent and her desires and dispositions to act. The descriptive version of the instrumental principle cannot, of course, sustain this distinction since it is just not possible for any agent to *conform* their own behaviour to a principle that is thought to be valuable only for giving third-person descriptions of actions. Consider some comments from Korsgaard for example:

> And if I am to constitute *myself* as the cause of an end, then I must be able to distinguish between *my* causing the end and some desire or impulse that is 'in me' or causing my body to act. I must be able to see *myself* as something that is distinct from any of my particular, first-order, impulses and motives....[If] [t]he distinction between my will and the operation of the desires and impulses in me does not exist,...that means that I, considered as an agent, do not exist.[67]

The picture painted here is clearly one of an inner deliberator who governs, controls and chooses how to act. Even if the ends are already given the agent has the ability to

[67] Korsgaard (1997), p.247.

choose the best means and decide which end should have precedence in any particular circumstance. A similar view is expressed by Nagel:

> Once I see myself as the subject of certain desires, as well as the occupant of an objective situation, I still have to decide what to do, and that will include deciding what justificatory weight to give to those desires.... One is suddenly in the position of judging what one ought to do, against the background of all one's desires and beliefs, in a way that does not merely flow from those desires and beliefs but operates on them.[68]

McDowell, in a similar vein rejects what we are calling the IR view on the basis of what he sees as the implausibility of a view that maintains a sharp distinction between desire (or dispositions) and cognitive states. He argues:

> It does not seem plausible that any purely natural fellow-feeling or benevolence, unmediated by the special ways of seeing situations...would issue in behaviour which exactly matched that of a charitable person; the objects of a purely natural benevolence could not be guaranteed to coincide in all cases with the good of others as a possessor of the virtue would conceive it. It seems still less plausible that virtuous behaviour in general could be duplicated by means of the outcomes of independently intelligible desires.[69]

McDowell questions and rejects the view that there always has to be an independently intelligible desire that action (for example, moral action) is trying to satisfy. The picture of agency being criticised here is one of a purely or merely (depending on what way you look at it) desiderative aspect entirely separate from and governed by, the practically rational, principle-following part – the part the agent identifies as *herself.* What she governs are her unruly desires and impulses that are capable of offering resistance to her will. The governing part of the agent (the agent herself) conforms to the instrumental principle as a matter of normative necessity. Taking the means to one's chosen ends is a constitutive norm of action in the same sense as swinging the golf club and attempting to hit the ball in the direction of the pin is a constitutive norm of golf. One just could not be playing golf unless this norm was followed. To say this

---

[68] Nagel (1997), pp.109-10.
[69] John McDowell (1978), p.20.

is *constitutive of* golf simply means that this is what golf consists in, that is, if this is not followed then one is just not playing golf at all. So conforming to the instrumental principle has to be, on this view, a constitutive norm of action in a similarly fundamental way. This means of course that we are supposed to reject the idea that the instrumental principle is only capable of providing descriptions of actions. So what we are left with here, according to this conception, is a fundamental norm that is constitutive of autonomous action. In so far as we are practically rational we just do have reason to pursue our ends.

If we accept this argument, however, there is one further step that it is claimed makes most sense to accept, namely, that our ends themselves must have some normative status. If an agent chooses to take the means to a given end, how can she maintain her focus on achieving the end in the face of distracting influences that might succeed in dissuading her? If the instrumental principle is to be taken as merely descriptive then all the agent can say is that she strongly desired that end at some earlier stage. But how can such a fact commit her to the same course of action *now*? Korsgaard suggests that the agent might succeed only if she is a 'heroic existentialist' who takes her *choice* of action to be normative and so can commit herself to achieving the end in question without considering the end itself *good*. Of course this suggestion has more than a hint of the ridiculous to it and we are supposed to reject it out of hand as one further step along the road to a complete rejection of the IR view.

Now I am not going to address 'heroic existentialism' head on because it seems clear that the crucial points of disagreement occur elsewhere. One of the principal issues I want to address is the rejection of the instrumental principle characterised as 'merely descriptive' and as such incapable of mounting a challenge to properly normative accounts of practical reason. If one can reject the instrumental principle of the IR view as uninformative and incoherent because it is a descriptive tool only, then

the above picture of the inner deliberator reveals as its companion another and similar rejection, this time of the possibility that dispositional states can capture what it is to endorse or commit oneself to norms of action. We are faced with the rejection of 'mere' description and 'mere' desires since neither can get us to the source of normativity.

### 3.5 'Mere' Descriptions?

From our discussion so far I think we can agree that descriptions alone cannot get us towards an understanding of what it is to endorse or commit oneself to a norm. However I do *not* agree that the naturalistic project (as 'naturalism' is commonly used in this debate) must now raise the white flag in surrender. The arguments against the IR view we have been examining do seem to present us with a heavy load of necessities but much of the force of these arguments can be shown to derive from contestable intuitions concerning the nature of agency. Understood as utilising a descriptive principle, the IR view is supposed to be incoherent. Now precisely the IR view or something like it is in fact shared by many contemporary followers of the broadly sentimentalist tradition,[70] but the attempts to give a naturalistic account of normative life in general tend to get dismissed for missing the unconditional requirements on practical reason and agency. It is argued that if there is an instrumental principle requiring us to take the means to our ends then there must be unconditional reasons for having certain ends. If we just happen to have certain ends then how can we be *required* to take the means to them? At most, or so the objection goes, we can simply *want* to take the means, but no requirement can be derived from merely wanting to take

[70] For example, see Alan Gibbard (1990).

the means to an end. Thus if the IR proponents try to defend the view that practical reason operates according to the instrumental principle alone and that this is sufficient to *require* the holders of ends to take the means to them then they must say something more than that the agent will take the means to her end because she *wants to*. But if they claim that the agent is required to act in a certain way then, we are told, the agent must have some reason to hold on to the end.

Imagine Mary today has buying a house as an end and then tomorrow she has the end of blowing all her savings on a sports car. If the instrumental principle is interpreted purely as saying that an agent takes the means to her end then Mary will have no reason to keep her savings safe to buy the house and if she decides that she wants to give all her money to charity five minutes before she goes to buy the car then she will have no reason to go through with that purchase either. If the IR theorist wants to argue that Mary is capable of *choosing* between incompatible ends rather than always taking the means to the most immediate end then there will have to be a way for Mary and agents like her to say that some of her ends are better than others.

However according to the objection, nothing like this can be derived from the instrumental principle. If the instrumental principle cannot function as the principle from which we derived *all* our reasons to act then the IR theorist has no option but to admit that there must be some unconditional reasons for having certain ends. If Mary decides that buying the house is what she should take the means to achieve then there must be some reason in favour of this that is unconditional, such as the value of independence she is committed to in being the owner of her own home. In short then, the objection is that if there are instrumental requirements on action (which all IR theorists want to be the case) then there must also be unconditional requirements on ends.

Korsgaard, for example, takes this even further. There are two options we face now at the point at which we must accept or reject the necessity of unconditional principles. We can either recognise that instrumental requirements are "self-given laws" which we give to ourselves because we are autonomous or self-governing, or, we can reject altogether the idea that there must be unconditional principles. If we choose the latter option however we cannot fall back upon a view of practical reason as purely (that is, descriptively) instrumental, because that position is supposed to be incoherent. So the rejection amounts to what she calls "...complete practical normative scepticism..."[71] and this is nothing other than a denial of *agency* itself: the agent becomes a passive playground for competing impulses, desires etc. Practical reason then, on the PPR view, is grounded in our autonomy and what we have to figure out is whether our autonomy determines the content of the unconditional principles of practical reason. Korsgaard of course argues that our autonomy *does* commit us to certain substantive principles.[72]

### 3.6 The IR Agent: A Jumble of Desires?

What I want to suggest is that the PPR view is right up to a point, but that nevertheless much of the argument presented above is directed at what is somewhat of a straw man. IR theorists are not generally committed to the view that practical reason can be constructed along crudely descriptive instrumental lines.[73] Most importantly, it seems more than a bit of a leap to argue that once it is accepted that the instrumental principle cannot stand alone the only option left open to us is some form of rationalism or Kantianism. Surely showing us the faults of a view which grounds practical reason

[71] Korsgaard (1997), p.253.
[72] We will be discussing these in the next chapter.
[73] Gibbard (1990); Blackburn (1993, 2000); Frankfurt (1988, 1999); Bratman (1987, 2003).

solely upon the instrumental principle does not thereby show that it must be unconditioned principles that give reasons their normative force. The PPR theorists are right in their diagnoses of the problems encountered by common versions[74] of the traditional Humean thesis. On a very basic level we know that sometimes we think certain things worth wanting irrespective of what we actually want and in a similar vein there are things we want that we do not consider to be worth wanting. We can also accept that describing nature is not what we do when we engage in normative talk. To this extent at least the PPR theorists are preaching to the converted. But what cannot be accepted, I argue, is the implicit claim that the only way to account for all sorts of action (and irrational action in particular) is to conceive of practical reason as governed by substantive principles from which an agent derives reasons through the exercise of her own autonomous willing, as conceived by the neo-Kantians we have been considering.

If an IR view can account for practical reason without reducing the agent to a jumble of impulses and desires and if we can plausibly account for all the relevant sorts of action then we will clearly not be inevitably pushed towards a rationalist view of practical reason, or as I have been calling it here the PPR view. It is more than plausible to argue, I want to show, that there can be at least one other working conception of practical reason and practical agency that *can* offer us an alternative picture of our normative life. There might still be a possibility, however, that 'irrational' is not a term which we can apply to actions with any great explanatory power. Now we saw the complaint made against this type of claim already – if we cannot account for irrational action then how can we have any action classed as *rational* (and therefore also considered to be right or required action)? What I will be

[74] I say this because I do not agree that Hume himself is the crude descriptivist he is sometimes supposed to be. However it would be somewhat tangential to pursue this any further.

suggesting later is that we might indeed have to let go of these terms for evaluating actions as they have been traditionally construed, but that in the broader arena of normative talk I think we can come up with an alternative characterisation of our fundamental evaluative framework for actions that captures a basic normative sense of approval and disapproval of the actions of ourselves and other agents. What we will ultimately need to pay a little more attention to is the issue of what it is to *accept* norms and have them guide our practical reasoning rather than concentrating first on the issue of what it is for something (such as a reason) to *be* rational or normative.[75]

The attempt to ground practical reason solely on the instrumental principle[76] is represented by PPR supporters as an attempt to place constraints on what could possibly count as a principle of practical reason. Because these constraints are allegedly justified by scepticism about the capacity of reasons to motivate agents to act through the mere recognition of reasons as considerations or *all-things-considered* reasons in favour of action, any successful argument showing that scepticism about the possibility of unconditional principles of practical reason is unjustified is supposed to show us that the IR view simply cannot account for action (rational or irrational) or practical reason more generally. The accusation is that the IR view, if it is to have any validity as an account of practical reason, needs to *derive* scepticism about the possibility of substantive unconditional principles from the claim that reasons alone cannot motivate. If it can be shown however that this derivation is impossible because the IR view relies on a view of practical reason as purely instrumental (along with the view that the source of all motivation lies in our dispositional states) then PPR proponents seem to believe that they have neutralised the threat to their view – the only

[75] Interestingly this is Hume's strategy for understanding causality. He argues that what we need to do is interpret our examination of the *nature* of causality, and look first at what is involved in making a causal *inference*. Hume argues that the latter inquiry answers the former. *Treatise*, I.iii.3.

[76] The objection is, as we have seen, that the instrumental principle cannot be converted into a normative principle that can act as *the* stand-alone principle of practical reason.

view which retains the possibility of giving us unconditional substantive principles that will inform all our practical reasoning.

As far as weakness of will is concerned, the IR view, if we accept the PPR picture of it, can provide no plausible explanation at all. If a distinction cannot be drawn between the agent's ends and what she actually pursues then anything the agent does is what she 'wants' to do and doing what one wants (or achieving one's ends) is the only requirement of the sole principle of practical reason – the instrumental principle. Because this distinction cannot be drawn then there is no way to account for irrational or wayward action – there could be nothing that an agent could do that was not the end she wishes to pursue at any given moment. There can be mistakes made, as we have seen Hume describe, but there can be no irrationality and so nothing worthy of being called a truly weak-willed action. But this is not a satisfactory conclusion according to PPR theorists. And since it is supposed to be the only conclusion the IR view can come to, the only option left to us is to reject entirely the IR view of practical reason as operating only according to the instrumental principle.

I believe we can seriously question these objections to the IR view. Having taken a detailed look at the PPR characterisation of the IR view what is becoming clear is that it is not the IR view in general that is in fact coming under attack. Instead what the PPR defenders seem to be focusing their criticism on is the 'no genuine judgement' version of the IR view we discussed in Chapter Two. As we saw there, according to the 'no genuine judgement' view, if an agent fails to do what she *believes* is better for her then all we can say is that she must have had a stronger desire to do the 'worse' in the first place and so therefore she really has an all-things-considered reason to do the worse. In other words, the agent was simply deceiving herself by thinking that she had endorsed doing the better as her all-things-considered reason for action. Somehow she was unaware of the true extent of her motivation to do the worse and so we must say

that the judgement or reason which she believed was her all-things-considered reason for action was not the reason on which she was motivated to act (and did act). If we take Hare's view as an example here, the main idea behind the 'no genuine judgement' view seems to be that it is not possible to make an evaluative or normative judgement and *not* recognise that this is what one ought to do; thus akratic agents cannot be said to make proper evaluative judgements in the first place.

In opposition to this view the PPR theorists themselves appear to defend what we might call the 'true irrationality' view. According to this view if an agent fails to do what she knows is better for her then she must be truly irrational in the sense that there must be some breakdown in her responsiveness to available reasons for action. If a reason is a *guide* to action then it seems clear that it is necessary to make a distinction between what an agent pursues and what she thinks she ought to pursue. If the distinction cannot be made it is difficult to see how any agent could be *guided* by normative principles at all. If the IR view can only show how agents can be motivated but cannot show how they can be guided then it does seem difficult to see how any theory of practical reason could be developed from this. If we equate the IR view with the 'no genuine judgement' view then it is difficult to see how the IR view could indeed be developed as a normative theory of practical reason.

However this is where one of our problems lie. PPR theorists seem to take it for granted that the IR view is identical with the 'no genuine judgement' view and consequently the prospects for an IR account of practical reason always come out looking rather bleak, if not hopeless. It goes almost without saying then that on this view of practical reason weakness of will can only have one 'solution' – to argue that it simply does not and cannot occur when the agent has considered her options and determined on a course of action.

But the IR view is not restricted to the 'no genuine judgement' view. Earlier on we saw that motivational internalism (both strong and weak) has certain deficiencies as an account of what it is to act on a reason or to have a reason for action. If we accept this conclusion (and I think we should) and also accept the PPR criticisms of the IR view (which I think we should *not*) then it is easy to see that we might be led to believe that there is only one direction left for us to go, and that is towards a PPR view of practical reason and action explanation. This is not an implausible suggestion. If the IR view cannot provide an account of the instrumental principle that is sufficiently normative to stand alone as the sole principle of practical reason, if there is no way to distinguish between what an agent pursues and what she holds as an end, then, it is difficult to see how anyone could continue to be interested in the IR view.

However the PPR view arguably ends up asking too high a price from us. The trouble lies in the PPR defender's insistence that one endorses a reason (or, takes it to be normative) when one evaluatively concludes that that reason is best, all-things-considered. This not only rules out many ordinary actions as irrational but begs the question with regard to weakness of will because it assumes that weak-willed actions are examples of motivational problems only in order to show that motivational problems pose no threat to the central notion in a theory of practical reason – endorsement of reasons for action. Yet before we tackle endorsement we should first of all tackle the objections to the IR view and expose their weaknesses. These are not sound objections and so this strategy, as a defence of the PPR view, seems unworkable. In the following chapter we will see why these objections do not work.

# Chapter Four

## Weakness of Will and Endorsing a Reason for Action: the PPR Criticisms Examined

At the end of the previous chapter we concluded that the real target of the PPR criticisms is only the 'no genuine judgement' view, a view which we found problematic for independent reasons earlier. Combined with our dissatisfaction with motivational internalism in its various existing guises it is tempting to accept the PPR arguments – especially if we find successful the argument that the IR view cannot provide an account of the instrumental principle that is sufficiently normative to stand alone as the sole principle of practical reason.

However despite their initial plausibility I think we can deny these claims. The argument I want to put forward here centres on the rejection of one claim. Negatively put it amounts to the denial of the following:

> That an IR view can amount to nothing more than a causal/descriptive thesis about how agents *do* act (and therefore unable to distinguish between the agent's ends and what she actually pursues).

When we have worked through the objections to this claim I hope it will be clear that the view the PPR advocates are in opposition to is a view which *is* implausible and unworkable, but most importantly, it is a position which no contemporary IR theorist should feel that they must, or have no choice but to occupy.

More positively, the sophisticated IR position I want to work further towards here recognises that practical rationality cannot be purely instrumental, recognises that

talk of 'what it makes sense to do' is fundamentally *normative* talk, and recognises that an account of practical reason that works only on the instrumental principle cannot capture our *acceptance* of norms that guide or govern our beliefs, moral judgements and actions. None of this, however, requires the formalism that, say, Korsgaard or Nagel[1] seem to think it does. Part of the problem we have to deal with here is the metaphysical argument put forward especially by Korsgaard, who claims that morality (or, rectitude in practical deliberation) derives from our nature as *deliberating* agents and so it cannot be rejected in favour of an IR perspective or any other view "...unless we are prepared to reject practical normativity, or the existence of practical reasons altogether."[2] If we can show that a sophisticated IR view can give an account of our deliberative nature that respects the above conditions then much of the strength of Korsgaard's rejection of it will have dissipated. To this end will be instructive to start by taking a look at Korsgaard's discussion of Hume's "sensible knave." If we can see this agent as properly deliberative[3] and if reflection on the knave's moral sentiments does not (contra Korsgaard) somehow undermine or weaken those sentiments, then it will be difficult to see why the PPR criticisms are considered unavoidable and devastating.

At this stage to continue to represent this emerging alternative position as an IR view (in the sense of a view whose only standard is the instrumental principle) would be misleading, and to distinguish it from the other views we have discussed I think the term 'Reflective Dispositionalism' (RD) better suits what I have in mind. As we shall see, on this view, what it is to endorse some consideration as a reason for action is to reach a stable intention or commitment regarding one's evaluative judgement. The

---

[1] At least according to Nagel's views as expressed in his 1970 *The Possibility of Altruism*. He seems to claw back from his conclusion that any genuine reason for action must be agent-neutral. See Nagel (1997).

[2] Korsgaard (1996), p.125.

[3] That is, having the capacity to assess one's inclinations and dispositions where these assessments are action-guiding and also guide *future* deliberation.

agent can then go on to evaluate her commitments in further practical reasoning and in this way her commitments become self-governing or self-directed rules for living. This account does not lack any essential notion of 'willed commitment' as Korsgaard argues that any broadly IR theorist view must do. We will be focusing in this chapter on evaluating the PPR objections to such a view and this will lay the groundwork for a more detailed discussion in the following chapter about how practical reasoning on this alternative account actually works.

### 4.1 The Free Rider

What we are trying to do is develop our understanding of what Kantians really mean by saying that a weak-willed agent is unresponsive to reason. To this end what we need to figure out is whether either Hume or the RD advocate can explain the action of an agent in a way that can give her reason for acting the status of somehow being a conclusion of reflection. That is, is there enough on a suitably modified IR (that is, RD) view to capture the *endorsement* of reasons for action?

Hume's original 'sensible knave' has a character we might describe as somewhat machiavellian. Whenever opportunity for personal gain presents itself he takes it, taking precaution only to ensure that his reputation is not sullied and that the essential fabric of society is not threatened with dissolution. It may be right and correct to conduct one's general commerce with humanity honestly at all times, but since there inevitably are exceptions even to this rule (as one might suppose) it would be prudent

to take full advantage of these.[4] Hume finds such an agent – a genuine sensible knave – difficult to convince otherwise. If he feels *no* natural reluctance to follow such a path then he has lost "a considerable motive to virtue" (*E*233). The virtuous agent, on the other hand, will have an "...antipathy to treachery and roguery [that] is too strong to be counter-balanced by any views of profit or pecuniary advantage." He continues: "Inward peace of mind, consciousness of integrity, a satisfactory review of our own conduct; these are circumstances, very requisite to happiness, and will be cherished and cultivated by every honest man who feels the importance of them." (*E*233) The problem with the sensible knave in relation to our concerns here is that if the instrumental principle is the sole principle of practical reason then he is entirely 'rational' in his knavery despite our obvious disapproval of him. But Hume sees the knave as more or less lost to morality and herein lies the source of a real and persistent rejection of Hume's view. To admit the possibility of hopeless knavery is interpreted as a weakness of the overall system – if the knave can get by (and probably do quite well for himself) then the rest of virtue just seems supererogatory or optional. The danger perceived is that morality is stripped of any recognisable *authority*, and the most Hume can do is recommend, exhort or chastise. But if he cannot point out to the knave that he is blind to authoritative principles that practical reason *must* operate under if it is to count as practical reason at all, then the fear is that this type of view amounts to no more than either a moral nihilism or irrationalism.

Korsgaard's version of the knave is an interesting modern update. A lawyer is faced with a situation where just as she is about to execute the will of a deceased client a new, superseding will turns up and in this one the client specifies that all his money is to go to his no-good nephew instead of medical research as he had originally stipulated. The lawyer considers whether she should ignore this new will and persist with the

[4] *Enquiry Concerning the Principles of Morals*, IX.2, pp.282-3.

earlier one since the money would be used for the general good instead of for funding the activities of a hell-raiser. But having read Hume the lawyer knows that she would disapprove of herself if she did this because it would be an offence against justice to do such a thing. As Korsgaard points out, the lawyer fully understands why she feels as she does. She knows that her disapproval concerns the consequences for the system of justice of the *general practice* of ignoring valid wills. But since no one would ever find out what she did this particular case would be exceptional in having no bad effects. Now as far as Korsgaard is concerned, "[i]t is almost inconceivable that believing this will have no effect on her disapproval itself."[5] Her criticism of Hume (and Williams) really amounts to what she sees as a misunderstanding of the purpose of reflective endorsement. Since the Humean lawyer sees her disapproval as being about the general practice of ignoring wills the way is open for her to see something different, something mitigating about this case which she can execute in complete secretive privacy. No one need know. The client is, quite literally, as silent as the grave, and therefore she has no reason to have a feeling of disapproval or self-loathing since this is a special case and not an example of a wicked sentiment caused by flouting a general rule. She now has a *reason* to ignore the original will and as far as Korsgaard is concerned she has given herself this reason by reflecting on and endorsing the sentiment in favour of ignoring the will she already has. In other words, she justifies what she already considers a sentiment worthy of moral approval. So we are back again to a picture of the agent as bundle of desires and dispositions where the guiding power of deliberation and reflection is reduced to rubber-stamping already formed sentiments.

There are two issues arising here. Firstly, we have to consider whether the IR theorist (if not Hume then the RD advocate) can explain the lawyer's action (of deceit) in a way that can give her reason for acting the status of somehow being a conclusion

[5] Korsgaard (1996), p.87.

of reflection (or some sentiment reflectively endorsed). If the IR theorist cannot do this then we may have to accept that there are insufficient grounds for normativity here, no way to explain an agent's internal conflict or do anything more than *describe* actions (rather than explaining their justifiability as part of our normative life). Secondly, even if the IR theorist could explain the lawyer's action in terms of her reflection on and endorsement of a general rule against unjust practices there is still the argument from the metaphysics of the self or moral identity. Her reflection and endorsement could never be the result of distancing herself from the desires and dispositions she must reflect upon and this is the crucial difference. For the PPR view personal autonomy is the source of obligation and the way this is reflected in action is through my own self-legislation. So for example, the lawyer requires, on this view, a principle in order to make a choice about how to act. It cannot just be any principle – it must be one of my own choosing or otherwise my ability to reflect on my desires and inclinations and choose how to act will be compromised by some principle whose origins I know nothing about. Now all this requires the assumption that I have a free will in a particularly strong sense of unconditional or non-dispositional ('transcendental') freedom, which is a significant consequence to be asked to accept. We will be coming back to the issue of what is required for an agent to have autonomy later, but it should be noted that there is no doubt that there is a high price being asked, with only a questionable transcendental argument as security.

As agents, and as a matter of personal identity we identify ourselves with whatever principle we choose to influence our choice of action. So on the PPR view the humean lawyer is far from autonomous. If I am required by my own self-legislation to do something then the reasons I give cannot be just my motives for acting. In other words, I *can* be motivated to do what I am obligated to do but my reasons for being

obligated are not identical with or supervenient on, my motives. All the lawyer has, according to Korsgaard, is a general rule or principle that she does not believe can be applied in her particular situation. Her qualms about what she is doing simply get in the way of acting on her other sentiment. She has no respect for the Moral Law and so she must, unfortunately for her and those whose lives she affects, be heteronomous and not a true agent at all. The IR theorist then, on the PPR view, cannot explain reflective, deliberative agency because she has a picture of agency where reflection on one's desires and concerns leads to the undermining of those desires and concerns rather than to the endorsement of some of them as reasons for action. Let us examine this issue next.

### 4.2 The Deliberative Agent

Central to the dispute between the two views we have been discussing is a disagreement over the nature of the *motive* of *obligation*. We all agree more or less that the practice of virtue somehow involves acting on the motive of obligation but whether there is any more to obligation than this is the matter of dispute. This issue arises under another guise – the concept of moral *rectitude* or, the capacity to act according to *principle*. Can we plausibly explain what it is to live the virtuous life without a prior account of obligation or rectitude? The IR tradition argues that we can explain obligation in terms of the virtuous action that is the result of stable dispositions and natural motives. Acting from the motive of obligation *just is* acting on the natural motives and stable dispositions one has. A properly disposed agent in this position

would feel a sense of disharmony or self-disapproval if she did otherwise. On this view there is no need for a further motive of duty, obligation or a special capacity to act from moral rectitude to practice virtue in a way that is morally inescapable for right-thinking agents.

What gives rise to significant uneasiness about this view is the apparent reduction of obligation to an entirely agent-relative conception of reasons for action. If one approves of a motive to act because it contributes to the virtuous life (which in turn amounts to the happy life) then the sense of obligation that emerges is one that is relative to the interests of the agent. So we could say that it is in our interests to be the kind of agents that practice virtue for its own sake. But the difficulty here is similar to the problem with the knave – how can we convince someone that she has no good reason to act, for example, from purely selfish motives without any consideration for others? Will the knowledge that others disapprove, that she will experience shame and self-loathing be enough to turn her to the virtuous path? Even if this could work with some agents the problem is that it provides no guarantee, no way to move the free-rider types who claim they simply do not care about the disapproval of others or who just will not experience shame, remorse, resentment or self-loathing.

On the PPR view the free-rider or knave can be easily picked out as having no good reasons for what they do. This can only be done, it is argued, if the right sort of metaphysical or quasi-transcendental[6] scaffolding is in place and this, of course, centres on the notion of moral identity. Korsgaard's argument for practical identity has three moves: (1) if we are to act at all then we require some sort of practical identity; (2) practical identity at its most fundamental involves being a reflective human being; (3) being a reflective human being entails valuing one's deliberative, reflective nature in a way that commits one to morality as a whole. Because we can reflect on our desires,

[6] Or at least a certain privileged status for the 'first-person perspective of agency and free-will' etc.

inclinations etc., we take a necessary *step back* from them and ask ourselves whether we can form a reason for action based on the desire under deliberation.[7] So, for example, if Mary has a strong inclination to spend lots of money every Saturday night then she must reflect on that inclination to discover whether she has a reason to act on it. It will be a reason for action if she finds that spending lots of money does not violate a principle which she must use to determine her choice of action. But this principle must be arrived at through her own reflection otherwise her reasons for action might not be her own. Mary cannot appeal to some principle she just happens to be familiar with such as "do not spend money you haven't got" since this is not something she has arrived at by reflecting on whether she herself could regard this principle as *binding*. So if Mary reflects and finds the principle "do not spend money you haven't got" binding but decides that she *does* have enough money to be able to splash out every Saturday night then the reason she gives for doing this cannot just be her inclination for doing so. When Mary reflects she finds inescapable the idea that she ought to do what on reflection she concludes she has reason to do.

What is noticeable about Mary is that she reflects *upon* her inclinations. It seems to be essential to this argument that reflection is something we engage in that is, in a sense, done from a position of neutrality or of complete detachment from one's dispositions. Examples abound. Consider for instance Korsgaard's claim that a clear distinction can be drawn between the 'acting self' and the 'thinking self'. She goes on to say that "[t]he acting self concedes to the thinking self its right to govern". Or consider the explicit claims we met briefly earlier:[8] the picture painted here is clearly one of an inner deliberator who governs, controls and chooses how to act. Even if the ends are already given, the agent has the ability to choose the best means and decide

[7] See Korsgaard (1996), p.104.
[8] Ch5, p.42.

which end should have precedence in any particular circumstance. Nagel finds himself "...in the position of judging what one ought to do, against the background of all one's desires and beliefs, in a way that does not merely flow from those desires and beliefs but operates on them."[9] And McDowell finds it "...still less plausible that virtuous behaviour in general could be duplicated by means of the outcomes of independently intelligible desires."[10] The most crucial assumption behind this thought is the claim that a rational practical reasoner (or, an agent with a rational *will*) can only act under the 'Idea of Freedom.'[11] This is supposed to be necessary for and perhaps even supply a sufficient guarantee of the morality of rational conduct. But does this mean that it is not possible to choose to do anything irrational? Some interpreters of Kant see agents as unable to will anything contrary to rational principles. On this view my will is always good and I always will to conform to the laws of autonomy. The implication of this, of course, is that anyone who wills to act, immorally or on a lesser reason for action all-things-considered must not really will to act at all. We have to say that they are not really free or they are heteronomous rather than autonomous.[12]

However I think we can find in Kant a more plausible PPR position than this interpretation would have us believe. 'Will' in Kant's *Groundwork* is a "kind of causality" that is distinct from causation by prior events and natural laws. It is instead, an ability to bring things about that is a unique feature of rational agents (*G*446). 'Will' amounts to the agent's capacity to act "in accordance with his idea of laws – that is, in accordance with principles" (*G*412, 427). To be a rational agent with a will one has to have the capacity to make things happen in such a way that the appropriate explanation

---

[9] Nagel (1997), pp.109-10.
[10] John McDowell and I.G. McFetridge 'Are Moral Requirements Hypothetical Imperatives?' *Proceedings of the Aristotelian Society*, 1978.
[11] To be fair to Kant, the 'Idea of Freedom' is something that he finds inevitable and not something that he *begins* his argument with. However this is not entirely clear. He begins chapter three of the *Groundwork* with the idea of freedom (when he is turning from the 'analysis' in chapters one and two to the 'grounds' of obligation); and in the *Second Critique* freedom is part of the 'fact of reason' from which he starts.
[12] See Robert Wolff (1973).

of the action is only gained through reference to the principles or reasons on which the agent knowingly acted. One implication of this view of the will is that Hume's *reductio* of indeterminism does not apply because Kant is not saying that the will is indeterministic. What Kant seems to want to deny altogether is the idea that there is a causal, empirical account of why free wills act for the reasons they do. In the *Critique of Pure Reason*[13] he claims to have established the possibility that our rational will can in some sense be responsible for what we do despite his proof that empirical science can in principle give us causal explanations for all mental and physical phenomena. But there is a high price to be paid if such claims are to be consistent – he has to conceive of the will as something *somehow* independent of the spatio-temporal causal order.

It appears, then, that Kant regards the free will as capable of acting independently of motivation by desires and inclinations. Most interestingly for our purposes however is that he seems to consider the agent's own desires and inclinations as 'alien determinations' of the will. The immoral agent has this kind of will. Contrary to the interpretation mentioned above according to which heteronomous actions are not freely chosen (in which case it becomes difficult to see how we could be *blamed* for our immoral actions), Kant holds that the agent who allows 'alien influences' to determine her will – and thus acts as if the 'instrumental principle' were the sole authoritative principle for her will – *freely* does so in the sense that she *could have done otherwise.* That is, she is the sort of agent who is acting on the wrong sort of principle of action (or 'maxim'). The agent acting on such a maxim ends up with a practical identity that takes as its exclusive authoritative principle the *instrumental principle,* and, as we saw earlier the argument is that this leads both the agent and ourselves as others trying to interpret her actions, further away from being able to differentiate between what the

[13] In the Third *Antinomy* of the *Critique of Pure Reason.*

agent considers her ends and what she actually does. This practical identity is ultimately incoherent, then, as the agent is left without any principle that can genuinely guide her practical reasoning and so insofar as she acts in this way she contributes to the development of a character in which she effectively has no practical identity at all and is more akin to a wanton, blown about by tempestuous desires and inclinations.

The crucial distinction then is between *being caused* to act by my inclinations and *choosing* to act on principles that make satisfaction of a particular inclination the reason for acting.[14] Acting autonomously is about a will's "being a law unto itself (independent of every property belonging to the objects of volition)" (*G*440). If rational agents are free in this sense then the next conclusion (in the *Groundwork,* at any rate), that such a will is autonomous in the sense of self-legislating the moral law, is reached via a transcendental argument. If there is free rational agency then there must be unconditional principles of conduct. Agents have *the capacity* insofar as they are rational to act on non-hypothetical principles or imperatives that they consider binding. Agents who fail to act on the principles they consider binding are experiencing irrational conflict of will. Of course Kant is not trying to deny that in a very broad (and therefore weak) sense agents *wanted* to do what they did do, for there is a trivial sense in which this is true. However it seems equally true to say that many of his opponents did not do what Kant in effect accuses them of doing, that is, of making every end depend upon the agent having some purely *self*-interested inclination or disposition towards it.[15]

It is important to point out here that Kant is not just talking about *seeing* oneself as free as this is something inescapable in rational deliberation. Rather what he has in mind is that a rational agent is one who takes full account in her deliberations of all the

[14] See Thomas Hill, jr., pp.257-8 'Kant's Argument for the Rationality of Moral Conduct', in Paul Guyer (ed.) (1998).

[15] In the *Groundwork*, chapter one Kant grants that someone might have entirely other-regarding feelings of humanity but he says that action done from these feelings is still not action done from duty.

implications of the assumption that she is free. Kant's strategy is to move from a conception of freedom in the negative sense (free from control by 'alien causes') to autonomy and then from autonomy he proceeds to demonstrate the necessity of morality. Any rational agent as such is committed to a principle which she acknowledges as rational but which is not based on any inclination to satisfy certain desires she may have. As a reflective, thinking being *who acts* the agent cannot but act on principles (or what a Humean might consider a 'general rule), and so there must be some way in which the agent *governs* her choice of action or justifies to herself the principles she considers to be rationally binding. At the most fundamental level Kant is attempting to demonstrate that as rational agents we need some such principle of choice.[16]

A moral agent qua acting morally, then, will choose to act on a rational principle she finds inescapable or binding and thus endorses. If she reflectively endorses a principle that would authorise a certain course of action then she is supposed to have already considered and reflected upon her "...incentives and impulses, including affections which arise in a natural way."[17] So no rational agent as such (that is, qua acting rationally) can be involved in the type of conflict in which desires and inclinations are meant to be opposed by 'reason'. A rational agent already has a commitment to a principle of action and when she reflects she cannot but think that she ought to do what she concludes she has reason to do. When she fails to do what she concludes she ought to do we are asked to accept that she is not experiencing *conflict* in the sense that she simply and brazenly fails to do what she knows she has already

[16] It is a matter of dispute whether Kant actually establishes that the Moral Law (on one formulation, act only on principles that all rational agents can will as universal law) is *the* principle that free rational agents must accept as rationally binding on any choice of action. Kant seems to say that the "supreme moral principle" is equivalent to the "principle of autonomy". In other words being autonomous is going to be the same as being a moral human being since acting on principles that all rational agents can will as universal law is equivalent to acting according to the principles of a rational, autonomous will. Whatever principles autonomous agents find rationally binding seem to be identified by Kant with moral principles.
[17] Korsgaard (1996), p.127.

concluded she should do. Instead she is "unresponsive to reason" and this is what true irrationality consists in. If an agent fails to do what she knows is better for her then she must be experiencing some breakdown in her responsiveness to available reasons.[18]

While there appears to be a similarity here with the 'no genuine judgement' view I think this view is actually more like the weak internalist solution. Korsgaard argues that we cannot infer from an agent failing to do what she concludes she should that she never really had the judgement or the reason to act in the first place. Rather because a truly autonomous agent wills to act on principles she considers binding we cannot ignore the 'willed commitment' of the agent when it comes to following through with action. The only explanation of problematic cases of conflict given these assumptions is supposed to be obvious – there must be a sense in which the agent is not fully rational.

Korsgaard's suggestion (following Kant) of a breakdown in responsiveness to reasons is intriguing. What this amounts to is something we have seen already when we discussed the weak internalist responses to the problem of weakness of will. The will is weak in the sense that the agent suffers what amounts to *practical failure* and it is anybody's guess as to what the cause is.[19] One thing however is supposed to be clear: qua rational and autonomous, failure to act as one thinks one should is not possible. Things can interfere with the internalism requirement – rage, passion, depression etc. – all these things could cause us to fail to be motivationally responsive to the reasons available to us. The point is, we are told, that reasons per se move us *necessarily*; it is of the nature of a reason to be sufficiently motivating. However this

---

[18] Korsgaard (1986).

[19] Practical failure, to recall, is any sort of psychological, emotional or other interference in the normal transitions in practical reasoning. So for example, someone feeling depressed might practically reason correctly and then not be motivated to act. It is not their capacity to practically reason that is supposed to be affected. Rather the argument is that they are not properly responsive to reason. So in a sense we can say that they are just experiencing motivational difficulties. Weak-willed agents are in this category, according to Korsgaard's Kantian view.

does not mean that reasons necessarily succeed in moving us to act. I can be irrational not just when I fail to see the rational connections (e.g. the failing to see that the sufficient means to the desired ends are available) but by also being wilfully blind to them (forms of self-deception come to mind here). The analogy with theoretical reason is supposed to make things clearer: many things could cause me to fail to be convinced by an argument even though I may be capable of performing logical operations. Failing to be convinced does not amount to any kind of evidence of a lack of ability.

Interesting as it seems, this view does not go very far towards an explanation of irrational action or internal conflict. Furthermore, as we saw earlier this strategy is dubious because of its inability really to tell us anything more about an irrational agent other than *that she is irrational* and that her irrationality need not remove from the account of practical reason under discussion the ability to categorise an agent as irrational.[20] What is troubling about this approach is that it leaves us describing such an agent as in some sense less than fully rational, and whereas this might not be as problematic as claiming that she must have had 'no genuine judgement' to begin with, it can be seen to be a step too far towards this type of explanation as will emerge below.

Let us take an example. Consider our old friend John who has to take a language test to get his much-desired job in the civil service. According to Korsgaard's argument John qua rational agent will reflect on his desire for the job and will act on it *if, and only if,* he endorses some principle according to which he can decide whether he *really* has a reason to take the test and get the job. What might that principle be? Because he is a rational agent with an autonomous will he will act on principles that he

[20] If a view of practical reason was not defeasible in this way then we would face a difficulty: an agent who fails to choose to do an action that will lead to the satisfaction of an end that she considers to be in her best interests and knows that this is what she is doing cannot be called irrational. She either must not have this as her end in the first place or doing what is in her best interests must not be important to her (i.e. she does not have the motive to act that, for example, externalists would attribute to her). But once we allow practical reason to be defeasible then we do not have to accept this analysis of a case such as this – we can say that such an agent simply is irrational.

can will as universal law. Now in his case that might be something like 'make all reasonable effort to find oneself a meaningful vocation in life'. So there John is, a citizen in the Kingdom of Ends, deciding that the best thing to do is to take the test like any similarly placed rational agent would do. But then he fails to take the test and can offer no exculpating circumstances.

Now we are supposed to say that John has suffered a breakdown in his ability to be "responsive to reasons." All that really springs to mind here is some kind of psychological or emotional disturbance. Because John *qua rational agent* would not choose this for himself (by his own admission the best thing to do is to take the test) then this assurance is supposed to relieve us of our worries over how he or any other agent could possibly act *irrationally*. Our problem is only a pseudo one if we accept the PPR view, as John is not *genuinely acting* irrationally because he is not acting autonomously in the first place.

Once again it seems that *defeasibility* forms the central pillar of the explanation of weak-willed and generally considered irrational action. The attraction of such an explanation is clear – the PPR view of practical reasoning can stand without threat from supposed counterexamples like weakness of will, which neither has to be denied nor the weak-willed action reconfigured as what the agent 'really' wanted or as rational relative to the agent's own beliefs. Weakness of will is just a defeasibility condition on practical reason according to this view and so poses no threat to the (PPR version of the)internalism requirement. Like the weak internalist solution we met earlier, Korsgaard's version of the PPR view seeks to avoid the problems that are the result of taking the strong internalist approaches to weakness of will.[21] But as we saw, the weak

[21] To recall, strong internalism is the claim that having a reason R to do A is logically sufficient for having the corresponding motivation to do A (and hence leads to the doing of A upon recognising R). So the connection between R and (the motivation to) A is necessary and indefeasible. Weakness of will is a case of inferential (logical) failure and so is irrational in the troubling sense that it becomes questionable whether we have really explained it. It appears that strong internalism must deny weakness of will but

internalist view is not without problems and in a similar way these problems re-emerge for the PPR view.

We saw in chapter two that in order to avoid the 'no genuine judgement' view a popular strategy involves weakening the connection between judgement and motivation to something more like a causal or contingent connection. The consequence of this, we saw, is that weakness of will becomes a case of practical rather than inferential failure. However on this view to say that an akratic agent is irrational is no longer an answer to the problem that holds any explanatory force; for to label an action as 'irrational' in the way this view does in no way explains what has gone wrong with the agent's practical reasoning. Now anyone who holds the PPR position will find themselves with a similar conclusion despite the fact that they have an entirely different view of the relationship between an agent's judgement and her motivation.[22] The reason why an agent fails to act as she thinks she should is not primarily about her rationality but must be something to do with weaknesses in her overall psychological makeup. Her failure to act then is very much a *practical* failure because she is as much 'unresponsive to reasons' on the on the PPR as she is construal weak internalist construal. The trouble is however, that we can get no further along than this. Despite their attempts it seems that PPR explanations of weakness of will face as much difficulty as those earlier views in trying to make consistent the two intuitions about practical reason we began with, that is, the internalism requirement and weakness of will.

The upshot seems to be that the practical reasoning of rational agents is inescapable except when it is not (i.e. when the agent just fails to be moved by her

---

the 'usurping reasons' view tries to preserve the two key intuitions about action (i.e. the internalism requirement and weakness of will) by arguing that the agent is motivated by a reason she has R, to do A, but not by her best or strongest reason. Defenders of the PPR view however, do not seem to consider a 'usurping reasons' solution to the problem at all.

[22] The PPR view of course does not view the relationship between judgement and motivation as a 'connection' that can be construed as contingent in various ways. As we have seen the PPR view is that the motivation to act necessarily emerges out of the agent's choice of action because the choice itself is the result of her reflection on her desires, reflection that is freely done using a principle that the agent herself considers inescapably obligatory.

reasons) – an occasional akratic breakdown in otherwise normal agents does not amount to evidence of the impossibility of reasons to act that are not based solely on the agent's desires or dispositions.[23] The view here of course is that if there are rational considerations that affect deliberation and choice (something like rationally binding or categorical imperatives) then it does not seem possible (for the *rational* agent) to go on to wonder whether she ought to follow them. In fact McDowell's view is in this respect more clear an example of how the PPR response to the problem of weakness of will ends rather unsatisfactorily. (PPR theorists do not all hold the same view here and the extent to which weakness of will is accepted as a defeasibility condition varies.) McDowell's views on weakness of will are similar to the 'no genuine judgement' view,[24] whereas Korsgaard comes closer to the practical failure view put forward by the weak internalists. No PPR defender (to my knowledge) attempts to do something along the lines of the 'usurping reasons' view because they all assume that weakness of will is just a motivational problem to begin with – something that would never afflict the *rational* agent qua rational. It is this assumption, I hope to eventually show, that causes the PPR view to beg the question regarding weakness of will and thus to fail to see the significance of the issue. First, however, it will be instructive to examine McDowell's treatment of weakness of will.

---

[23] Or, as McDowell puts it, it seems right to "...agree that one need not manifest irrationality in failing to see that one has reason to act as morality requires..." But he claims we can deny the conclusion he believes IR theorists derive from this, and that is "...that moral requirements are only hypothetical imperatives." (McDowell (1978), p.13.)

[24] Of course categorising McDowell as a PPR theorist might confuse some into believing that he must be more Kantian than anything else. However this is not the impression I wish to convey. McDowell prefers to situate himself in what seems to be a middle ground between Hume and Kant by defending a view that sees the requirements of morality as being imposed not by some metaphysically necessary 'ought' or 'should' but by the way agents see the circumstances of their prospective actions. In other words we can locate the morality of the action not in the agent's desires (so in this sense his view will be non-hypothetical) but in certain features of her conception of the situation she is in (and in this sense it will be non-categorical). Whether McDowell's view is indeed the *sui generis* position he seems to believe it to be is quite another matter and beyond the scope of our present discussion. However insofar as he argues that desire need not and should not function as an independent component in the explanation of an agent's action we can put his view under the PPR heading for our purposes.

On McDowell's view a full explanation of a reason must explain how the reason was capable of motivating the agent to act; otherwise it could not be a reason at all (i.e. if it's not what made the agent act then it cannot be a reason at all). Furthermore "[i]t must contain enough to reveal the favourable light in which the agent saw his projected action."[25] We tend to assume that revealing the agent's *desires* are sufficient to show this favourable light. But according to McDowell "...it seems to be false that the motivating power of all reasons derives from their including desires."[26] For example, we explain some agent's action by crediting her with awareness of some fact which makes it likely that acting in a certain way will be conducive to her interests. Describing the situation in terms of the agent's view of the facts may suffice on its own to show us the favourable light in which the action appeared to her. But McDowell agrees with Nagel that ascribing a desire to the agent is not a necessary component in the explanation of her action. Just because an agent is motivated by a particular consideration (for example, a desire for her own happiness) does not mean that her reason for action can be reduced to an account of its motivational efficacy. It is not necessary that we have both a belief/judgement and a desire as independent components to explain an action since somehow the judgement/belief is motivational itself:

> If we credit him with a suitable desire, then, as before, that need be no more than a consequence of the fact that we take his conception of the circumstances to have been his reason for acting as he did; the desire need not function as an independent component in the explanation, needed in order to account for the capacity of the cited reason to influence the agent's will.[27]

McDowell points out an objection here. What if another person had the same conception of the situation but did not see herself as having any reason to act? McDowell gets around the objection by arguing that the virtuous agent allows of no

[25] McDowell (1978) p.15.
[26] ibid.
[27] ibid., p.16.

defeasibility – if the agent has the same conception of the situation but does not think she has a reason to act then she just does not really have the same conception of the situation:

> But if one concedes that a conception of the facts can constitute the whole of a reason for prudent behaviour, one is not at liberty to object to the very idea that a view of how things are might not need supplementing with a desire in order to reveal the favourable light in which someone saw some action. [28]

So we do not need to have a desire as a component of our practical reasoning to have a reason to act. When we act prudently we see the situation in a particular way and this view of the facts of the matter (e.g. what would be most sensible to do) *just is* my reason for action and this in turn *just is* the favourable light in which I see the action. I act and I do not need any other desire in order to push me to act. If this is true in the case of prudent behaviour then there does not appear to be any significant factors preventing us from applying the same analysis to virtuous behaviour. This is Nagel's argument of course, and McDowell seems to fully endorse it. To clarify:

> It is not that the two people share a certain neutral conception of the facts, but differ in that one, but not the other, has an independent desire as well, which combines with that neutral conception of the facts to cast a favourable light on his acting in a certain way. The desire is ascribable to the prudent person simply in recognition of the fact that his conception of the likely effects of his action on his own future by itself casts a favourable light on his acting as he does. So the admitted difference in respect of desire should be explicable, like the difference in respect of action, in terms of a more fundamental difference in respect of how they conceive the facts.[29]

If we take an example of a weak-willed action we will find that the PPR view seems to be able to provide us with only the same sort of explanations we met when we discussed the motivational internalists. John fails to sit the exam that he needs to pass so he can get the job he really wants. Now according to the PPR view, on the IR view the only criticism that can be made of John is that he might have made mistakes about

---

[28] ibid., p.16.
[29] Ibid., p.17.

the means to his ends. But there is no perspective of practical reason on the IR view from which we can accuse him of not acting as he thinks he should.

Unfortunately for the PPR view, however, the options for explaining John's failure to do that which he thinks he has best reason to do do not seem much different. Korsgaard will say that he is unresponsive to reasons[30] and McDowell will say that he must never really have had the conception of the situation that would make recognising that he has a reason to act inescapable for him.[31]

Now this seems terribly close to the framework through which motivational internalists invite us to view irrational action – either as inferential failure or as practical failure. The trouble is that as we have already seen neither of these options seems to work. If the IR view can be rejected on the grounds that it is a theory of practical reason that cannot satisfactorily explain common cases of action such as weak-willed action, then I propose that we can reject the PPR view on similar grounds. Now PPR theorists might respond to this that while they use what we have characterised as strong internalist (the 'no genuine judgement' view) and weak

---

[30] To recall the difficulties weak internalists get themselves into bear in mind what it means to be 'unresponsive to reasons'. Maybe we can isolate the problem with such an agent (say, John above) – perhaps he feels under stress, has no self-confidence or maybe he is just lazy. Whatever problem John might have the point is that the door is opened to the vast regions of psychological explanations for akrasia and all peculiar behaviour. On this view to say that an akratic agent is irrational is no longer an answer to the problem that holds any explanatory force, for to label an action as 'irrational' in no way explains what has gone wrong with the agent's practical reasoning.

[31] Now this might strike us as not being the conclusion we expected from a PPR view. If the connection between having a judgement or reason for action and the motivation to act is a necessary connection then the reason why an agent fails to act as she thinks she should might very well be about her rationality. Instead of following this line of argument however the popular view is that it must be something to do with weaknesses in her overall psychological makeup. Her failure to act then is very much a *practical* failure, not some sort of logical failure analogous to a failure of theoretical reasoning. While it may be an interesting debate as to whether the PPR advocate is entitled *only* to explain weakness of will as some sort of failure of *reasoning* (i.e. inferential failure of some kind *other than* the 'no genuine judgement' view) given the view of practical reason that it presents I am not going to pursue this question here. Let us grant the PPR advocates their views that supposedly 'irrational' action is explained either by some version of the 'no genuine judgement' view or as practical failure. These views are not necessarily contradictory. For example one might argue that the agent in question must not have had the correct judgement in the first place because of some pre-existing condition such as anger, laziness, short-sightedness etc. that prevented her from engaging in a practical reasoning procedure that results in a proper reason for action. However it is sufficient for our purposes to suppose that PPR advocates tend to hold to *either* the 'no genuine judgement' view *or* the practical failure view. A position combining both of these does not amount to a substantively different explanation.

internalist strategies for explaining weakness of will, this does not entitle us to reject their view of practical reason. They believe that they have defused the threat this poses and therefore they can use these strategies without suffering the same pitfalls because these are good strategies for dealing with motivational problems (i.e. they are consistent with the internalism requirement rather than counterexamples to it). However if these strategies still fail to satisfactorily explain weakness of will then we need to figure out why this is. In this chapter we will see how the PPR explanations of weakness of will fail. The reason why they fail is that weakness of will is conceived of as a motivational problem only. But to believe this is to beg the most important questions regarding weakness of will. Before we take a look at this however, I think it is important that we broaden the discussion somewhat.

What we need to take a closer look at is deliberative *experience*. The PPR view argues that a certain kind of deliberation – reflective endorsement – is something which agents are necessarily or logically committed to. The IR view, it is argued, cannot capture this. Or, to put it another way, it cannot explain the self-legislative (self-governing) aspect of rational agency. Because there can be no distinction between an agent's ends (or what she considers herself to have reason to do) and what she actually pursues the IR view cannot explain the kind of self-government that is the defining characteristic of rational, deliberative agency. We might say then that it is because there cannot be a distinction between the rational and the irrational or weak-willed agent that the IR view fails as an account of practical agency. Now what we should examine is whether we really need to argue that we are necessarily committed to reflective endorsement in order to explain the inescapability of deliberation (reflective endorsement) in rational agency.

Korsgaard argues that no 'motivational considerations' (such as weakness of

will) can weaken the logically necessary commitment to deliberation and reflective endorsement. But is this really true? I think there are reasonable doubts about this, and one way to raise these doubts is to take a look at deliberative experience as some of the 'motivational skeptics' (as Korsgaard calls them) might see it. If these 'skeptics' can give us an account of practical deliberation, derived from deliberative experience, that relies on a regulative or reliable norm of reflective endorsement (as opposed to arguing that we are logically bound to this as the guiding principle of practical agency) then perhaps motivational considerations really do place "constraints on what counts as a principle of practical reason."[32] On top of this, however, I will argue in the following chapter that these are *not* just 'motivational considerations' (to call them this begs the question). What is ultimately important is that a sophisticated IR view (RD) *can* talk about normative considerations (regulative ones) as guiding practical reason in a way that both meets the demands of deliberative experience and offers a more satisfactory account of irrational action.

### 4.3 Deliberative Experience

What the PPR view tries to get us to accept is that any rational agent qua rational has to be committed to a principle by which she chooses how to act, a principle that she finds inescapable for her, or, in Kantian terms, *a law*. The distinguishing feature of these principles is that they enable us to choose to act on some consideration *other than* our inclinations, desires, dispositions or anything that could be considered to be derived purely from experience. But what kind of principles are these? As we have seen

[32] Korsgaard, (1986) p.8.

Korsgaard argue for example, these are the principles which we can regard as binding on all agents similarly placed. Of course choosing a principle that one could consider a law is nothing other than the exercise of autonomy, and so practical reason (and of course morality) is a matter of following the principles an autonomous agent endorses. Now the principles that autonomous agents accept are supposed to be rational principles that can be accepted independently of desire or of any other factor placing constraints on one's ability to choose principles of action.

Rational agents then are supposed to conform to the universal principle of autonomy, but this principle has been defined so as to exclude rational considerations that are based on inclinations or desires. Because desires are defined as external influencers of choice we can see that this is something that autonomous agents must accept. However one point of difficulty here is that there is what seems to be an illegitimate move from the logical commitments of autonomy to morality. The problem that arises is this. Reflectively endorsing certain considerations is by itself supposed to commit agents to thinking that this reflective endorsement alone constitutes a reason for action. However unless we are able to share a conception of identity then this is clearly consistent with the possibility that agents follow the correct reason-forming procedure but might come to radically different and opposed conclusions about which actions are moral or permitted by practical reason. We all think of ourselves as autonomous agents and to this extent we share the same concept of agency, but we may all answer differently when asked what *sort* of autonomous agent we are. If I see myself as a teacher then the maxims I endorse will be limited to deciding how to act from this perspective. It is not difficult to see how seeing myself as a student might result in my endorsing quite different maxims indeed.

Korsgaard attempts to establish that we have a more fundamental identity that is

common to all perspectives on our own agency and that is our 'moral identity' as human beings ("a reflective animal who needs reasons to act and live"[33]). When we take things to be important to us then this must be because we take ourselves to be important, and so we can say that we value our humanity. Practical reason on this view is conceived of as a practice which is mandatory or inescapable because it is reflectively compelling. It is a practice that is in response to ought-judgements, that is, judgements made by agents that are prior to any other demands such as institutional morality or social mores. To put it another way, morality is the "...practice of 'oughting' ourselves and others into doing the socially cooperative thing."[34]

However we have to consider the relationship between the *form* of the ought-judgement or reason the agent comes up with and the content or substance of the reason. What is crucially important is the role of deliberation in all of this. Judgements about what I ought to do require, oblige or compel me to act but it is my *deliberation* that provides the content of my reason for action. Deciding how I ought to act is the practice of endorsing whatever my deliberation or reflection finds compelling. What makes an action something I ought to do is not, strictly speaking, because *I* can will as a law the principle that makes actions like this permissible, but it is because my deliberation considers the action to be worth choosing.

Let us imagine how such a rational deliberator might conduct herself. We can assume that she stops to carefully and clearly consider the consequences of her actions and she recognises that these implications will affect the decision she makes about her action despite whatever other desires or inclinations she might have. She will also recognise that she has endorsed or committed herself to exercising correct deliberation

---

[33] Korsgaard (1996), p.121. Some interesting commentary on the arguments in Korsgaard (1996) can be found in Alan Gibbard 'Morality as Consistency in Living' *Ethics*, 1999; Michael Bratman, 'Search for the Source', *The Philosophical Quarterly*, 1999; Rachel Cohon, 'The Roots of Reasons', *Philosophical Review*, 2000.

[34] W.D. Falk, 'Morality, Form and Content', in Falk (1986), p.235. This is a seminal discussion of many of these issues.

and will try to always make it the case that her deliberation determines her actions. This means that she must acknowledge the results of her deliberation and not give in to the temptation to evade the clear awareness of what ought to be done – a temptation made all the stronger by her likes, dislikes, fears, habits etc. It could be said here that such an agent cannot go through the ordinary tasks of life thus ponderously deciding that *this* is what she ought to do on any given occasion. It is more plausible to say that the moral agent has good habits and inclinations that lead her to act as she knows she ought to. She is at the ready, we might say, to reflect on any situation she finds herself in, and will not shirk the conclusion she must reach due to her own reflective endorsement of certain principles of action. So the self-legislating agent does not fail to do anything she judges that she ought to do because she is able to resist the urge to do the contrary. This is what we would consider a morally strong agent. But the PPR view does need to accept the view that the self-legislating or autonomous agent develops in themselves dispositions to act that, as we might say, *pass the test of reflection*. This broader conception of the autonomous agent really is necessary if the PPR view is to have plausibility. Well let us grant them that.

The next question we have to ask is whether the only way to make sure that one's dispositions pass the test of reflection is to argue that they must be sanctioned by reason. I want to argue that this is not the only option, and furthermore, that we can get the same sense of 'being moral' from a sophisticated IR view. To pre-empt the obvious objection, however, the burden on any alternative account is to explain something that neither view wants to deny: the inescapability of a commitment to reflective endorsement or deliberation. What the PPR view gives us is an account of morality that comes directly from what it is to be an autonomous agent. Our self-legislation is not something that we can take or leave, it is who we *are* and so how

could an account with a different starting point achieve the same categoricity? To preface the response: there is one claim here that is shared with the view I want to defend, and that is that deciding how best to act is an inescapably deliberative and norm-governed process. It may come as a surprise to some PPR defenders but contemporary IR thinkers are as committed to the inescapability of a self-governing process of deciding how best to act as those who hold the PPR view. The disputes are best categorised as about whether or not these IR accounts *succeed* rather than about the fundamental commitment to a shared and inescapable process of reflective endorsement.

In the next chapter I want to discuss the nuts and bolts of deliberative process involved in deciding how best to act and this discussion will focus on the norm-governed nature of this process. An agent can, I hope we will see, come to have a commitment regarding how to act that is reached by reflecting upon all relevant considerations, and these include desires, dispositions, previous intentions etc. The weak-willed agent's actions and practical reasoning can be understood not in terms of a defeasibility condition (that is, by introducing conditions that pose no threat to the internalism requirement) but rather as examples of common sorts of actions that should cause us to reconsider the extent to which we should accept that reaching a conclusion of reflection (or, an evaluative judgement) is, in itself, sufficiently motivational. I argue that it is not sufficiently motivational and that a more adequate conception of practical reason must recognise instead that there are further *norm-governed* transitions in practical reasoning that must take place *between* reaching a conclusion and acting on it.

Those that favour the PPR view seem to preclude themselves from this type of solution to the problem of weakness of will. As we have seen, however, the

explanations PPR proponents tend to actually favour are problematic in much the same way that the strong and weak internalists' explanations are (and we saw this in chapter two). For the remainder of this chapter I want to elaborate on two issues that have been mentioned here. I have suggested above that PPR proponents beg the question regarding weakness of will and now I want to be more explicit about this. Related to this is the problem of underdetermination of intention.[35] I argue here that it is when the problem of underdetermination is not recognised that one ends up begging the question regarding weakness of will. Let us take a look at this second problem first.

We have seen earlier how weakness of will is treated by a PPR theorist such as Korsgaard. Occasional akratic breakdowns in otherwise normal agents cannot be used to place constraints on what might count as a principle of practical reason, since "...the Kantian supposes that there are operations of practical reason which yield conclusions about actions and which do not involve discerning relations between passions (or any pre-existing sources of motivation) and those actions."[36] But the problem of underdetermination raises questions about our understanding of these operations of practical reason in the rational agent. Consider an example. Jim and John both independently decide that they will attend the lecture. To say that they *decide* to do this is to say that they both have evaluated the considerations available to them and have concluded that this is what they should do. This then is an evaluative judgement or conclusion of reflection. But only John actually goes to the lecture. He is somehow able to move from the evaluative judgement to the intentional action. In Korsgaard's terms he has been *responsive to (practical) reason*. In McDowell's terms Jim, on the other hand, must not really have had the evaluative judgement about attending the lecture in the first place. All underdetermination cases have this same form: of two

[35] In citing this problem my views are related to those of Michael Bratman on this issue. See Bratman (1987; 2003).
[36] Korsgaard, (1986) p.8.

agents who arrive at the evaluative judgement only one goes on to perform the act in question.

So what does John have that Jim lacks? It is fairly clear that the PPR view only has the resources to explain these cases as examples of weakness of will, or to put it generally, as examples of irrational action. It just does not seem possible for a rational agent to fail to act when her decision is based on rationally binding considerations. John seems to have an evaluative judgement that he finds inescapable and this (inescapability) is what his endorsement of his judgement consists in. However it is difficult to accept that this adequately explains the example, for presumably Jim too finds his evaluative judgement inescapable. Why should we say that he does not? All we really know is that he did not attend the lecture.[37] Have we, as the PPR view would seem to suggest, exhausted all possible explanations when we say that he is either unresponsive to reason or that he must therefore not have had the judgement to begin with? This seems rather limited in scope. I for one am not satisfied with such an explanation and would not be so willing to say that Jim's failure to act can be attributed to his weakness of will.

Let us leave aside the 'no genuine judgement' response to Jim's situation. It is just too much to accept that an agent who can evaluate his reasons and other considerations and reach a judgement based upon this reflection must not really have done any proper evaluation at all. We have dealt with this elsewhere. Let us consider the charge of being unresponsive to reason. By this I assume that what is meant (or what Korsgaard means) is that the agent is unable to respond in an appropriate manner to the conclusion of his reflection or his evaluative conclusion (let us use this latter term from now on). The appropriate response to an evaluative conclusion is, we can suppose, to endorse that conclusion and an agent demonstrates endorsement of this kind

[37] Of course the usual all-things-considered and ceteris paribus presumptions are at work here.

by going on to act on the evaluative judgement. So we can say that Jim does have the evaluative judgement but he is unable to act on it and show that he has endorsed it. This, at least, is how we are supposed to understand him on the PPR view (Korsgaard's version).

If we are unwilling to immediately judge Jim to be irrational or weak-willed then how are we to explain his inaction? To get an answer to this it is important not to ignore the other agent, John. Rather than asking what is wrong with Jim let us ask instead why John acts on his evaluative judgement. Is it because it is directly motivational? Now some theorists (such as Nagel[38]) might agree that it is. But instead of engaging here issues concerning the denial of the notion of unmediated desires, let us say on behalf of the PPR view not that the evaluative conclusion is directly motivational but that an agent who reflects and reaches an evaluative conclusion is necessarily committed to that conclusion and so will act on it, all-things-considered. Let us say that the agent is motivated to act on the *commitment* that necessarily accompanies the evaluative judgement.[39] This much should be acceptable to someone such as Korsgaard.

What is emerging from our discussion of the PPR view is that when an agent commits herself upon reflecting and reaching a conclusion, she necessarily commits herself to that conclusion (i.e. she commits herself to acting on her evaluative judgement). What the problem of underdetermination raises however is just this point. Why must we presume that an agent's commitment to act is determined solely by her evaluative judgement? Jim has the same evaluative judgement as John and yet he fails to have the same commitment (and we know this because he fails to act). I suggest that we cannot say at this point whether or not Jim is weak-willed. All we know is that he

[38] Or at least Nagel as in Nagel (1970).
[39] This commitment can be, but does not have to be, understood as a favourable attitude regarding the evaluative judgement.

fails to act because he does not have the commitment to his evaluative judgement that should motivate him to act. Since John has the evaluative judgement and goes on to act accordingly it might be more plausible to assume that he has a commitment to his evaluative judgement that perhaps is not derived purely from that judgement. Those who accept the PPR view and some of the internalists we discussed in chapters one and two think it is correct to assume instead that a commitment to evaluative judgement *is* derived purely from that judgement. They find it straightforward therefore to argue that the agent who fails to act is weak-willed. The weak-willed agent does not have the proper commitment either because he never really had the evaluative judgement to begin with or because he is unable to respond to his judgement in a way that allows him to recognise that he must have the necessary commitment or that he should form the commitment (depending on the view).

However I think we can question whether Jim's case really amounts to a case of weakness of will at all. In the following chapter I argue that weakness of will can be explained as difficulty in forming and sustaining one's commitments to action. The underdetermination problem brings us to issues concerning commitment rather nicely since it makes us question whether an agent such as Jim above really is weak-willed. We cannot tell from the information given whether he is weak-willed. It may be the case that Jim goes on to perform some action other than the one he evaluatively judges he should perform. The question we have to ask is whether it is always irrational for an agent to do so, and if so whether this amounts to failing to act as one thinks one should. This is the question for the next chapter where I argue that it is *not* always irrational to evaluatively conclude that *A* is best and end up doing *B* instead (after having judged that considerations in favour of *A* outweigh those in favour of *B*).

Before we get to that argument however it is important to point out and question

the assumption that an agent's commitment to act is determined solely by her evaluative judgement. If Jim has the evaluative judgement to the effect that going to the lecture is what he should do yet he fails because he lacks the commitment, what could it be that allows the other agent, John, to make the commitment to and then act on his judgement? I have suggested above that the commitment John has to his evaluative judgement is perhaps not derived entirely from that judgement. But what does this mean? In the process of figuring out what to do John reflects on his desires, attitudes, dispositions and other considerations and concludes that he should attend the lecture. So far John and Jim are the same. But John goes on to act, as we know. Here is the suggestion. John acts not because he knows that, when he reflects upon his reasons for action and endorses (i.e. he evaluatively concludes) one as his best reason all-things-considered, he will act accordingly insofar as he is rational. Rather he acts because he forms a commitment to acting that is reflected in his disposition to act (or we could also say it is reflected in his attitude favouring acting).[40] John has an attitude in favour of attending the lecture or a disposition to so act and he acts on this attitude or disposition. In the following chapter I argue that an agent's commitment constitutes her intention to act and that we must distinguish between an intention and an evaluative judgement.

In Korsgaard's terms, what we are suggesting is that the endorsement of a reason for action is constituted by the agent's commitment rather than her evaluative judgement. All I want to suggest at the moment, however, is that agents *do* act upon their commitments and that commitments are not entirely constituted by evaluative judgements.

The underdetermination case brings this to the fore. It is not at all obvious that the agent who fails to act is weak-willed since he matches the rational agent in all

[40] We could talk about commitment in terms of other cognitions.

respects that concern their practical reasoning. Whatever Jim is doing wrong the mistake is not to be found in his practical reasoning about the considerations for and against attending the lecture. It seems more plausible to explain the differences between them by suggesting that there is still a role in practical reason and action explanation for a notion of commitment that goes beyond the straightforward evaluative judgement. The PPR view has real trouble with the types of cases we have been discussing here because they make it difficult to justify the presumption that an agent's commitment to act is determined solely by, or indeed is constituted by, her evaluative judgement. It is more plausible to suggest that John's commitment to act goes beyond the evaluative conclusion he has that says "all-things-considered I should attend the lecture." Jim shares the same conclusion and there is undoubtedly something wrong with him, but again it seems more plausible to suggest that what this might be has more to do with his apparent inability to really commit himself to the course of action set out in his evaluative conclusion than it has to do with his practical reasoning. In a sense, then, Korsgaard's claim that such an agent is unresponsive to reason hits on the truth in a very broad sense but it is misleadingly vague when coupled with the implied suggestion that psychologists can fill in the gaps now that the philosophical work has been done. Surely it is philosophically interesting if the best suggestion for Jim's case is that an intention to act is underdetermined by evaluative conclusions about how to act.

There are further cases that seem to support this view that are not prima facie cases of weakness of will but are cases where rational agents have evaluative judgements that appear to underdetermine what finally turns out to be their intention to act.[41] For example, we can easily imagine a situation where an agent cannot reach an

[41] Michael Bratman has been examining such cases for many years now – at least since his criticism of Davidson on intending. See Bratman (1979) and (1999), pp.209-24. See also Alfred Mele (1986; 1987); Richard Holton (1999); Gary Watson (1977).

evaluative judgement of the form 'X is best, all-things-considered' because the alternatives (X and Y) are equally matched. Yet a decision has to be made. In order to allow for circumstances of this form it is more plausible to argue that the agent's commitment to act can go beyond the evaluative judgement (which in this sort of case will not be of the strongly evaluative form 'X is best').

If we continue to insist that the commitment to act is constituted by the evaluative judgement then we will have to find a way to resolve some difficult issues about intention. For example, in the case above if Korsgaard thinks that an evaluative judgement has the form 'X is best all-things-considered'[42] then we would have to say that the agent endorses doing X *and* she endorses doing Y because she has two evaluative judgements: she judges that doing X is better than anything else and she also judges that doing Y is better than anything else. But she cannot do this because she *knows* that she cannot do both X and Y. But then how is the agent to act? If she just chooses then she can be accused of being impetuous or of acting on a consideration that she has not endorsed as her reason to act. In cases like this the agent endorses some consideration as her reason for action, but not because she thinks that 'X is best, all-things-considered.' Perhaps she simply thinks 'X is just as good as Y.' In this case the agent cannot endorse doing X as the best option, all-things-considered and so in Korsgaard's terms she would not be properly endorsing the consideration in favour of X as a reason to act at all. This seems to place the bar too high for many of our ordinary reasons to count as properly endorsed reasons for action. Therein lies the implausibility of this suggestion.

If we adopt the PPR suggestion that the agent's commitment to action (her endorsement of a reason) is thus constituted by the evaluative judgement then these sorts of problems seem unavoidable. It also must be pointed out that the sort of case in

[42] Where 'X is best' means 'X is absolutely better than anything else.'

the example above is not even a case we would consider to be about weakness of will – it is an ordinary situation as any seemingly rational agent would experience it. And it does seem that in these ordinary sorts of cases the evaluative conclusion underdetermines the commitment to action the agent ends up with.

If the commitment to action goes beyond the evaluative judgement then what is it? I have suggested that this commitment is reflected in an agent's disposition to act.[43] But what sorts of dispositions are we talking about here? It would seem right to retain something of the sense of *endorsement* that the PPR view (and Korsgaard particularly) tries to capture. A commitment cannot be solely a favourable attitude towards whatever we evaluatively conclude we should do. If that was the case then any other fleeting attitude that happened to be stronger would probably end up as the dominant attitude of the moment. To avoid this we must say that a commitment must be a *stable* favourable attitude.[44] It must persist over time. Let us say I have considered all the factors I can think of in my current practical reasoning and I conclude that I should work on Saturday. I am committed to doing this if I have a stable favourable attitude towards working on Saturday (or, a disposition to work on Saturday that persists over time). This can be seen for example, in my refusing to consider an invitation from

---

[43] In this I am influenced by Bratman's views on intention. See Bratman (1999), pp.209-24 and (1987), pp.111-38.

[44] It is possible here to argue that commitment is not an attitude at all. Some might prefer to discuss this in terms of some other sort of cognitive state (perhaps more *cognitive* than attitudinal). I'm not sure about this but it seems initially plausible anyway. If the main point here is that evaluative judgements or conclusions (as I have been calling them) underdetermine the agent's commitment to act (or intention to act) then what completes this gap in the composition of a commitment is an attitude the agent has regarding her evaluative conclusion. I argue that that attitude must have certain features if it is to form an integral part of the agent's commitment to action. So what we are really saying here in a nutshell is that a commitment to act is a further mental state that is not identical with the conclusion of the agent's practical reasoning nor is it a simple attitude in favour of that conclusion. Like Bratman I argue that a commitment is not reducible to a piece of practical reasoning nor a simple attitude. However there is the possibility here to accept the underdetermination problem and argue instead that the further mental state that constitutes commitment is a cognitive state more like belief and that this belief must have certain features such as stability. My concern here however is to argue that there really is an underdetermination problem (and thus a begging of the question regarding weakness of will). I will consider my argument successful if this point is accepted and my suggestion about an *attitude* of commitment considered plausible.

friends regarding Saturday as having greater weight in my practical reasoning on say, Thursday.

Furthermore, my taking my evaluative judgement 'work on Saturday' to have greater weight in other practical reasoning (e.g. about what to do on Friday) tells us something else about my attitude. It tells us that I take myself to have good reason to work on Saturday. If we consider that I will persist in my commitment all through the decisions of the week ahead (declining other invitations etc.) then this commitment to work on Saturday must mean more to me merely than the idea that this is the plan for the weekend that I really wanted yesterday. There must be a sense of justification to my attitude as well as its stability. Once I have a stable commitment I will be firm and non-sceptical about it. In other words, it will take some very weighty and serious new information to get me to alter my commitment. In my practical reasoning about how to act on earlier days in the week my commitment about Saturday is authoritative for me. I am committed to working on Saturday and so this is a major consideration that will have the most significant weight in my further deliberations about the weekend. The stability of this commitment gives me a sense of its authority or its normative force for me. This is the sense of justification we speak of. This is reflected in my dispositions regarding further deliberation. I do not have any disposition to reconsider either my commitment or the practical reasoning and evaluative conclusion that lead me to make this commitment. This is what allows me to immediately refuse the invitation for Friday night. I do not need to go back and review everything all over again with this new consideration. Now obviously if I received a piece of information telling me that a family member is gravely ill and requires my assistance on Saturday then I will reconsider my position. This does not mean that my commitment is unstable or even that it disappears. Rather my evaluation of the situation has changed and my important

and overriding commitment to family becomes dominant. This commitment we might think of as a 'family first' policy lying dormant in the range of concerns that would influence my practical reasoning.

It is natural to think of agents as having certain values to which they are committed. We might understand this to mean that they have at some point in their lives they have practically reasoned and evaluatively concluded that something is to be considered most important in all relevant deliberation. They thus form commitments based on these evaluative conclusions and these commitments become influential in further practical reasoning. Family loyalty is one example here. Most agents would reconsider a stable commitment if it was inconsistent with a stable commitment regarding close family members. The other overridden commitment might still remain (depending on the circumstances) and even reassert itself once the other commitment has be acted upon. For example if my family member wants my help on Sunday instead of Saturday my previous commitment to work on Saturday will probably be reinstated.

This example is not quite like the underdetermination example above and yet it seems plausible for us to suggest that commitments go beyond the evaluative conclusion in these cases also. However we do not need to agree on this. Even if someone agrees that commitments (intentions) to act go beyond evaluative conclusions in underdetermined cases only then they will in effect be admitting that at least in some cases there is more to a commitment to act than an evaluative conclusion. Getting back to our agent, Jim, who fails to attend the lecture despite concluding that he should, can we now tell whether or not he is weak-willed? I think we can only conclude that we still cannot tell from the information so far supplied. For all we know of him is that he failed to attend. We do not know whether he went on to have the commitment to

attend, as the other agent did. He might have concluded he should attend and then simply not have had the corresponding commitment (intention) to do so. Or, alternatively he might indeed have made the commitment and then failed to intentionally act on it. These are the two possibilities for error on his part.[45] In the following chapter I shall be suggesting that we see both these errors as paradigm cases of weakness of will.

Let us turn now to the first problem mentioned in this section. We said at the outset that those holding the PPR view end up begging question regarding weakness of will because of their view that one endorses a reason for action only when one evaluatively concludes that that reason is best, all-things-considered. We are now in a better position to understand what this means. If we believe that what it means to endorse a reason is that an agent has reached an evaluative conclusion and will act on it, then the agent who fails to act on it must have a problem either reaching the right evaluative conclusion in the first place or else there must be something preventing her acting according to her evaluative conclusion. What that might be is either that she is being externally prevented from acting (i.e. she is a properly rational agent being prevented from acting against her will) or she is suffering some sort of psychological problem that leaves her listless, careless or in denial when faced with an evaluative conclusion about how to act. To say that weakness of will is something that prevents us acting rationally or that it causes us to be unresponsive to our reasons is to say that it is a problem about our motivation when faced with our own reasons.

---

[45] Of course it is still possible that Jim might not have made any error at all. He could have committed himself to going to the lecture but at the last moment circumstances might have arisen that caused him to give greater weight to a previous commitment (such as his commitment to his family). By this I do not mean that he necessarily practically deliberated all over again with the new information for it might be the case that he hears the news and rushes off without any time for practical deliberation of any significance. In this case we could say that his commitment to his family is a stable commitment (perhaps one that has come about as a result of deliberation earlier in life) that the agent himself treats as an overriding authoritatively guiding commitment. Very few circumstances could cause an agent to override such a commitment and this adds further to its stability.

However to say this is to adopt a question-begging strategy since this assumes the very thing we want to examine here, and that is how a *rational* agent practically reasons. To say that rational agents endorse a reason for action when they reach an evaluative conclusion about how best to act is the assumption that causes us to see weakness of will as purely a motivational problem. But this strategy fails to allow us to see that weakness of will really does cause us to question how *rational* agents practically reason and act. Perhaps part of the difficulty here is that our understanding of weakness of will as failing to act as one accepts or knows one should (given one's reasons) is simply a description of the problem that is too often mistaken for a sufficient explanation. But we cannot really start out by assuming that weakness of will is a motivational problem. From its bare description we do not know what sort of problem it is. At most all we can say is that it is an apparent or prima facie motivational problem only. The underdetermination case helps us to see this. If two agents have the same evaluative conclusion (in exactly the same circumstances) but only one of them acts on it are we really justified in saying that the one who fails to act is unmotivated either because of a psychological problem or because she is irrational? As we saw from the example we discussed we really cannot jump to either of these explanations.

What is interesting about the underdetermination case is that both agents appear to be rational (at least to the extent that they each can reach an evaluative conclusion about how best to act). We really do not have any grounds to say that the agent who fails to act is *therefore* irrational or weak-willed without further assumptions about rational practical reasoning. So, if both agents are rational (even if we only admit that they are so up to the point of reaching the evaluative conclusion) then a closer look is needed at how this is possible. Those who hold the PPR view would try to argue (I

assume) that the agent who fails to act therefore has a problem – a breakdown in the relation between her rational practical reasoning and her acting. However this claim is only available on the assumption that endorsing a reason for action guarantees that an agent will act accordingly (provided she is free to act, etc.) insofar as she is rational. It is this assumption that I examine and reject in the following chapter.

# Part III

In the previous chapter the suggestion was made that a sophisticated IR view could indeed account for practical reasoning or deliberation concerning reasons for action in a way that leads to the endorsement of a reason for action. The PPR view would have us believe that this is not possible for any IR view. However by outlining an alternative view of how practical reasoning actually works I suggest here that we can both satisfactorily account for weakness of will and avoid certain difficulties that a PPR account of practical reasoning leads us into.

Part III consists of Chapter Five and Conclusion. Chapter Five contains four sections. Section 1 outlines certain remaining internalist assumptions that I argue are still causing trouble for views on practical reason. Section 2 discusses the problems these assumptions cause for our understanding not only of weak-willed acts but of ordinary or enkratic acts. Any view which holds onto implausibly strong internalist assumptions ends up describing some cases of intending to act in a clearly unsatisfactory manner. Section 3 details a solution to these problems by highlighting a distinction between what is intentionally done and what is intended. I argue that there is a *normative gap* between what an agent evaluatively concludes about her proposed action and what she finally intends to do. Forming an intention on the one hand involves *commitment*, and acting on an intention involves *stability* on the other. Section 4 examines the implications of this alternative view of what is involved in practical reasoning. One important implication is that following the correct reasoning procedure as suggested by the PPR view cannot be guaranteed to bring a rational agent to act. The stable commitments required to fill the normative gap are not ones an agent can be rationally required to have but they are nevertheless crucial to agency. I argue

that the norms crucial to rational practical reasoning require certain dispositions on the part of the agent, dispositions without which rational action (moral or non-moral) will not occur no matter how well-reasoned the agent's evaluation of her situation is.

# Chapter Five

## Practical Reasoning and Intending to Act

### 5.1 More Internalist Assumptions

The PPR account of practical reason we have just been examining argues that weakness of will (as described) is impossible insofar as an agent is rational. As we now know this means effectively that weakness of will cannot on this view be described as acting (with full awareness) contrary to what one endorses all-things-considered. On this view endorsement of one's best reason for action is identified with willing or intending to act. Thus it does indeed seem that weakness of will as originally described cannot be possible.

What I want to argue here is that weakness of will *is* possible even though it is sub-optimal practice on the part of the agent. This argument does not rely on the redescription of weakness of will so we may accept the phenomenon as acting with full awareness contrary to what one endorses all-things-considered. Let us also accept that practical reasoning involves acting on an intention. However the assumption that endorsing a reason for action is intending to act is the claim that requires scrutiny.

There is one crucial assumption that leads us to hold that endorsing a reason all-things-considered is intending to act and that assumption can be expressed as follows:

> **A1**: Having an all-things-considered reason in favour of a course of action is necessarily motivating (all things being equal).

This can also be expressed in terms of endorsing one's reasons:

> **A2**: If an agent is rational then when she reflects upon her reasons for action and endorses one as her reason for action all-things-considered she will act accordingly.

'All-things-considered' in this case means that the reason is consistent with the agent's beliefs, with the general principles she accepts and with her conception of her practical identity as a Kantian such as Korsgaard understands it.

However A1 and A2 tend to camouflage or presuppose a more specific claim being made about the relation between having an intention and intentional action. This claim can be expressed as follows:

**A3**: If an agent intentionally acts then she must have intended so to act.[1]

Having a reason for action or endorsing a reason for action must motivate the agent to act since it just does not seem possible for an agent to intentionally act in a way she has not intended to act. There seems to be a tight or internal connection between intending and intentionally acting or between having an all-things-considered reason and being motivated to act on it. Up to this point we have examined this relation in terms of the latter connection but we cannot complete our study of practical reason and weakness of will without an examination of the former connection, which turns out to be crucial for understanding the processes of practical reasoning.

In arguing against A3 I will be obviously setting myself against A1 and A2 as well. However arguing against A3 does not imply that we must reject the basic idea that practical reasoning involves considering our reasons for action and coming to a conclusion based on our considerations. This cannot plausibly be rejected and as we saw earlier even an externalist can accept this. All they need to add is some external motivator that is not already considered or evaluated in the practical reasoning. So an agent can come to a conclusion about how to act and she will act because she has, say,

[1] Michael Bratman terms this the 'assumption of tight fit'. See Bratman (1987), chapter 8 and in Mele (ed.) (1997), p.194.

a general disposition to be motivated by her conclusions. What the rejection of A3 does involve however is questioning the assumption that the conclusion we reach when we consider our reasons for action is always about what we think is *best to do*. So if an agent comes to the conclusion of her practical reasoning that A-ing is best (all-things-considered) then according to A1 and A2 the agent will be motivated to act and will act. According to A3 there is no room for discrepancy between the agent's intention to act and her intentional action. So long as she is rational and so long as she has an intention she will (ceteris paribus) intentionally so act. There is no room here for weakness of will since this would involve the agent coming to a conclusion that A-ing is best, intending to A and then B-ing instead. What we would have to admit here is that her action is not intentional – she did not act in accordance with her intention to A. So it seems impossible that any agent could act akratically – that is, *intentionally* B when she intends to A.

In the context of A3 the ordinary rational agent's practical reasoning is fairly straightforward. Consider again Jane who has to decide whether to go for coffee now or to stay at her desk working. If she is rational Jane will evaluate all her reasons for both options[2] and then come to a conclusion about which reason she will endorse, and that will be her *best reason* for action all-things-considered. She decides to remain and work and so we say that she *intends* to stay at her desk. Because she is rational this is indeed what she does. However what is unsatisfactory about this picture of practical reasoning is the assumption that one's best reason is simply taken to be what one intends to do. Now at the very least this is a bit of a leap. It seems that there are plenty of apparent counter-examples – we can easily imagine Jane being weak-willed by deciding to remain at her desk but then getting up and going for coffee anyway. The

[2] Of course by saying that Jane evaluates 'all' her reasons we do not mean that she must consider all *possible* reasons she might have. Rather Jane evaluates all the reasons she can think of at the moment given the situation she finds herself in – a situation which might include the necessity for a quick decision, inability to remember certain facts of the case and all such ordinary limitations.

dieter who 'gives in' to the dessert trolley is another example. Now are we to say that these cases are impossible? What this would mean is that neither Jane nor the dieter act intentionally because they act against their intention to act. This seems like an implausibly strong claim to be committed to. An involuntary twitch of my leg muscle seems non-intentional, as does falling off the chair while reaching for a book. But Jane's (weak-willed) act and the dieter's act do not appear to be of the same type as these other sorts.

However if we accept A3 then we must say that either Jane's going for coffee is not an intentional action at all or that it is just not possible (i.e. the case is misdescribed) given that the intention to act was already formed. The possibility that has not been examined however is that Jane's action is irrational or sub-optimal but nevertheless intentional despite the fact that she has a different intention from the one that would normally result in her going for coffee. If this turns out to be the best explanation for these sorts of cases then our account of practical reasoning will have to change to accommodate this. The rejection of A3 would mean the rejection of a model of practical reason that has been around at least since Aristotle's claim[3] that the conclusion of a practical syllogism is an action. On this view, once an agent concludes her practical reasoning or endorses a reason for action then she acts and her non-action

---

[3] There is a debate between those who believe that Aristotle does allow 'clear-eyed' or true akrasia and those who maintain that he does not. Clear-eyed akrasia is what can be considered as 'calm' akrasia, that is, akratic action *not* performed under the influence of some disruptive appetite or other. In my view akrasia for Aristotle is to be understood as action performed as a result of an appetitive disruption and the resulting suspension of particular knowledge. Calm or clear-eyed akrasia just seems too incredible to be considered seriously (1146b35, 1147a9). Acting against one's knowledge or judgement when one has reasoned correctly remains a mystery — if indeed he really believes that it ever happens. Such situations are "extraordinary" or "astounding", but I believe that Aristotle is speaking of situations which could only exist in principle. *Nicomachean Ethics,* trans. Terence Irwin (Indianaoplis: Hackett, 1985). See Anthony Kenny, 'The Practical Syllogism and Incontinence', *Phronesis*, 11, 1966. David Charles, '*Rationality* and Irrationality' *Aristotelian Society Proceedings*, 83 1982/3. David Pears, *Motivated* Irrationality (Oxford: Clarendon, 1984). JJ Walsh, *Aristotle's concept of Moral Weakness,* (New York: Columbia University Press, 1963) E Anscombe, 'Thought and Action in Aristotle', in R Bambrough (ed.) *New Essays on Plato and Aristotle* (London: Routledge, 1965).
David Wiggins, 'Weakness of Will, commensurability and the objects of deliberation and desire', *Aristotlelian Society Proceedings,*76, 1975/6

can only be explained by reference to being prevented by external forces, impetuousness, self-deception and weakness of will (in the sense of 'giving in').

Is it possible then for an agent to intentionally *B* despite having reached an intention to *A*? What we shall see as we examine A3 more closely is that this is indeed possible. We will also need to consider whether this is always strictly irrational on the agent's part or whether there can be situations where it is rational to form an intention to act and end up intentionally doing something else. *Ex hypothesi* these cases would not be examples of weakness of will since there would not be any breakdown in the agent's practical reasoning. Weak-willed cases could only be ones where the agent irrationally *B*s despite intending to *A*. If this distinction can be made then we also need to say what it is that makes one case of intentionally *B*-ing while intending to *A* rational and another case irrational. What I hope to show is that this distinction *can* be upheld and that what makes one case rational and another irrational concerns the 'normative gap' between one's evaluative reasoning and one's intention formation. I argue that it is not an agent's reason or rationality that demands her to move from her evaluation of her reasons to the formation of an intention but rather it is her proper functioning as an agent according to a complex network of norms that she herself accepts, but that are not also logically required in practical reasoning, that guarantees the transition (and *guarantee* might even be too strong). Before we discuss these issues however we first need to establish whether indeed an agent can intentionally *B* despite having reached an intention to *A*.

**5.2 Must one intentionally act as one intends?**

We have seen quite a few examples of actions typically considered to be weak-willed that appear to contradict A3 above. The cases of Jane and the dieter above are such examples, as seems to be the agent we met in earlier chapters who wants to pass the civil service exams but knowingly fails to do the essential preparation despite intending to do so. In all the cases we have met it would seem to be rather difficult to persist in maintaining that each of these agents either fails to act intentionally or never really had the intention to act that they claimed they had. The best description of such agents seems to be that they intentionally *B* despite intending to *A*.

The claim that such agents are just not rational may well be true, but it amounts to a rather insubstantial explanation. The problem is that a bald response of this sort really does not tell us very much about the case of the *rational* agent either since A3 on most accounts is taken as true but defeasible – the agent who fails A3 must be irrational. However if the present proposal is correct and an agent can be rational and yet also fail A3, then any account relying on this type of internalist assumption will not be giving us a necessary condition (even if a defeasible one) on practical reasoning. If there can be cases of rational practical reasoning where it is *not* true that if the agent intentionally *A*s then she must have intended to *A*, then A3 will not express a necessary condition on practical reasoning. If this is the case then we will have to look elsewhere for necessary conditions. The weak-willed agent is important here because it is by examining the difference between rational and irrational cases of intentionally *B*-ing despite intending to *A* that questions concerning the necessary conditions of practical reasoning and agency may be more clearly brought to light.

So much for promissory notes, but are there cases where it might be rational to intentionally B despite intending to A? Let us take a look at some examples. Michael

Bratman[4] discusses an interesting case which is worth taking a look at here. The case concerns two video games and the goal for each game is the same – to hit the target. The agent decides to play both games simultaneously even though she knows that it is impossible to hit both targets since the games are so linked that if both targets are about to be hit the machine will simply shut down. Let us suppose that the agent believes that these games are very difficult by themselves and that even if the machine did not automatically shut down it would be extremely difficult for anyone to hit both targets since they would need excellent motor control in both hands (a rare enough skill).

Now the agent decides to give both games a try simultaneously since there is such a large prize involved if she succeeds that it is worth risking shutting the machine down. Lastly, let us suppose she hits target 1 (T1) and fails to hit target 2 (T2). The question we have to consider is whether she *intended* to hit both of the targets. It seems clear that she must have intentionally hit T1 since she must have put a lot of effort and skill into the attempt. According to A3 if she intentionally hit T1 then she must also have intended to hit it (this could also be expressed as intending to *try* to hit T1). Since she must have put equal skill and effort into trying to hit T2 then according to A3 she must also have intended to hit (or to try to hit[5]) T2. But the agent knew before she began that she could not hit both targets so how could she intend to do so?

This brings us to what some consider to be a condition or constraint on intention and for this I will follow Bratman's terminology by calling this the 'strong belief requirement' on intention. Davidson's view for example, implies that our agent playing the games would only have the intention to hit T2 if she did not have a belief with the content 'hitting T1 and T2 simultaneously is impossible'.[6] Since the agent in our

---

[4] 'Two Faces of Intention' in Mele (ed.) (1997), pp.183-5.

[5] We have to assume that A3 includes intentionally trying to act. Failure to do something cannot automatically be attributed to failure to intend to do that act.

[6] Davidson argues that intentions are unconditional (i.e. I do not have an intention to A *only if* B) – my beliefs act as assumptions without which I would not have an intention in the first place. So my intention

example knows (or believes) that hitting both is impossible she will not intend to hit T2. But does she intend to hit T1? Well according to A3 intentionally hitting T1 implies having the intention to hit it but this does not seem to do justice to the case described. An agent in such a situation would not just intend to hit T1 and at the same time try to hit T2 without having any intentions regarding it. The agent is in the same mental state with regard to hitting both targets so we would assume that she has the same intention regarding T2 as she has regarding T1. But these intentions fail the strong belief requirement test since she knows she cannot hit both targets and this means that she really could not have had those intentions to begin with. However as Bratman points out, all the agent is doing here is having a try at hitting *something* – she is just doubling her chances and this seems like a rational strategy given her circumstances. It does not seem correct to say that she has no intention to act because she believes she cannot hit both targets.[7]

The difficulty here involves a confusion between the conditions we believe intentional actions must fulfil and the conditions we might well want to place on having an intention. For example, Mark sees white paint blotches on his new red car and starts scrubbing. We could say that he intends to remove the marks and according to the

---

is conditioned by my beliefs but not *conditional* upon them. Intentions are (and must be) consistent with an agent's beliefs. Intentions assume, but do not contain a reference to, a certain view of the future. A present intention about the future goes something like this: 'given what I now know and believe, here is my judgement about what kind of action is desirable.' The intention is *based on* my estimation of the situation. Consequently it would be very inaccurate to say that I intend to act only if my estimate is right. If I believed that I could not act, or should not act, or I am not right in my appraisal of the situation *then I will not have the intention*. 'Intending' in Davidson (1980).

[7] It might be objected here that the case can be described as one where the agent is simply intending to 'try to hit one or the other' while knowing that she cannot hit both. In this way the agent has an intention which is strongly consistent with her beliefs about the games. But this restricts the situation to a specific description. The case outlined is one where the agent is trying to hit each of two targets and each target guides her perception of it. Her trying ends when she hits T1 – she wants to hit it, uses her skill to do so and stops when she perceives that she has hit it. She also tries to hit T2 and so we have to assume that she is in the same mental state regarding it except that she does not perceive that she has been successful. It just does not seem right to say that she has intentionally hit one of the two – it does not match the intentional states she was in during the attempt. (See Bratman (1987) p.118) Now the weakness of this example is that it seems all too readily to invite subtle redescription, thereby making the strong belief requirement seem plausible. But to redescribe this case is to change it, not to defend it. So the case I'm dealing with is one where the agent tries to hit T1 and tries to hit T2.

strong belief requirement this implies that he believes he will remove them.[8] But while scrubbing (intentionally acting) must Mark also believe he will remove the marks? Is it not possible that while acting he has worries and doubts whether scrubbing will work despite his intention to remove the marks *by scrubbing*? It seems quite plausible for a rational agent in this position to decide that scrubbing hard is the best course of action, to then intend to scrub, but while acting not have the belief that scrubbing removes paint from cars. Why would we want to say that Mark must believe in the general efficacy of scrubbing while engaged in the act? This same question arises for intending to act. John might intend to call Jane this evening and this is what he plans to do but he knows about all the things that might prevent him calling and so if asked he may say that he does not believe he will call Jane, but he intends to nevertheless. He is not saying that he firmly believes he *will not* call, rather he just does not believe or is not sure that he will. Given these types of examples Davidson's claim about the necessity of belief (i.e. the strong belief requirement) does not seem essential either for intention or intentional action.

If I can have an intention to act despite the belief that my attempt will fail and if we can intentionally act despite not believing that we will so act then we have grounds for rejecting the strong belief requirement. But belief must be somehow related to intending and acting, so what sort of relation is it? What makes John's case above seem rational is that he does not have a belief that is inconsistent with his intention to act. Not believing or not being sure that he will call Jane is not inconsistent with his belief that he will call. What would be inconsistent here is having the belief that he will call, the intention to call and then the belief that he will not call. This does not seem rational for any agent. Now someone could object that John is not rational despite not

[8] But by saying this we are not committed to the claim that intending to X is identical to believing one will X. What we are saying is that intending to X necessarily involves believing one will X.

having inconsistent beliefs and I think we have to agree that there is something criticisable about John's intending to call while being so unsure about whether he will actually do it. Although I do not think we can go as far as Bratman here and argue that John is criticisably irrational. Maybe he is culpable for entertaining beliefs that are *nearly* contradictory (there surely is something sub-optimal about an agent who sails as close to the wind as this) but it does not seem necessary that we consequently accept that there are logical errors in his practical reasoning.

I agree then that the strong belief requirement is implausible but I also agree that we should replace it with a consistency condition.[9] The claim can be expressed as follows:

> Consistency Condition: if an agent intends to *A* she will not have any beliefs that contradict beliefs she has about her intentionally *A*-ing.

Mark would fail to adhere to this requirement if he intended to scrub the car, had the belief that he would do so but believed also that he would not do so (perhaps because he believed he would bring it to be re-sprayed). Assuming that he knows he could not both scrub the car himself *and* bring it to be re-sprayed (because of time constraints for example), he clearly falls short of the consistency required by holding these contradictory beliefs. In this situation we can accuse Mark of being irrational and we might go on to wonder whether he really has the intention to scrub the car at all.

So believing that I will act is not a necessary condition of either intentionally acting or intending to act – I do not need this belief *in order* to intend or intentionally act. All I need then is to have a consistent package of intentions and beliefs. By extension we can see that this requirement is one that intentional action must meet

---

[9] Bratman argues for a consistency requirement on agents' intentions and plans. See 'Two Faces of Intention', p.186 and also (1987) pp.31-3.

also.[10] Mark intentionally scrubs the car so long as he does not have beliefs about his scrubbing that are inconsistent with his belief about what he is intentionally doing. So if he believes that he is driving in Monte Carlo this is inconsistent with his belief that he is acting on his intention to scrub the car. Now it might concern someone here that the requirement is not strong enough. The fear is that we could imagine a scenario where one could have a belief about what one is doing that is compatible with (but distinct from) what one thinks one is intentionally doing. Imagine Mark intending to scrub the car but believing that he is waxing it – these two are compatible since some time-saving products do both. However in response we can say that whereas the *activities* in question may be compatible the corresponding *beliefs*, taken individually, are not. The *belief* that I am waxing the car is not as it stands consistent with the belief that I am scrubbing the car. The only way these could be made consistent is by having a further belief that for example, scrubbing with Scrubowax cleans and waxes simultaneously. So where this belief is absent we can say that Mark does not meet the consistency requirement and we can then doubt whether he intentionally scrubs the car.

If we accept this consistency condition on intending and intentionally acting then we have to take another look at the practical reasoning of those agents who do not intentionally act as they intend. Can these agents be rational? The agent in the video game example makes an interesting case. She intentionally hit T1 but failed to hit T2. A3 and the strong belief requirement suggest both that she had the intention to hit T1 (because she this is what she intentionally did) and yet that she must not have had an intention to hit either T1 or T2 (because she knows that it is impossible even as she picks up a 'gun' in each hand). Anyone who wants to remain committed to A3 has to

[10] Bratman argues that constraints placed on intention do not necessarily apply to intentional action. The fear is that imposing a consistency requirement of the type under discussion on intentional action does not allow for the observation that one can intentionally act while doubting that one will. I argue however that my acting is intentional so long as I do not have beliefs that are inconsistent with what I believe I am doing. This does not mean that every time I intentionally act I must believe that I will.

accept that it gets its force from the strong belief requirement. So the agent can only have the intention to hit both targets so long as she does not have a belief that this is impossible. Since she does have such a belief she cannot have the intention to hit the targets and therefore when she hits T1 she is not acting on any intention regarding it. But according to A3 this is what she does so she must have had the intention to so act.

There does not seem to be any easy way out of this difficulty. A3 and the strong belief requirement go hand in hand and yet in this case they imply that the agent did and did not have an intention to act. Given what she knew she could not have had the intention and since she did act she must have had the intention.[11] None of these difficulties would arise of course if the agent was described as intending first of all to hit T1 and only on her second chance intending to hit T2. But this is not the case as described and this is where A3 and the strong belief requirement have difficulty.

What is really going on in this case is that the agent is doubling her chances of winning and to this end she intends to hit T1 and intends to hit T2[12] despite her beliefs. She intentionally hits T1 but does not intend to hit it any more than she intends to hit T2. I think what we have here is a case of intentionally *B*-ing despite intending to *A*. However we can only accommodate this commonsense description by rejecting A3 in favour of the Consistency Condition on intending and intentional action. What we can say is that the agent knows she cannot hit both, that she only gets one chance, that there is a nice prize to be won and so she has intentions to hit the targets. She has no beliefs that are inconsistent with any of this. But she does not believe that she *will* hit the targets and yet she still intends to try to hit each one. There is no inconsistency here because none of her beliefs is incompatible with either intending to try to hit each target

[11] Remember also that she knew she could not hit T1 and T2 simultaneously so she could not have formed the intention to try to hit both at the same time.
[12] We can read these also as 'intends to try to hit'. See fn 5.

or with intentionally hitting T1. She intentionally hits T1 despite intending to try to hit T1 and intending to try to hit T2.[13]

In this case we can say that the agent is not irrational in any way even though her intentional action and her intention do not share the same content. She does not fail the consistency requirement because her intentions and her beliefs are not in opposition. A3 would have us accept that the attitude the agent has when she intentionally acts is the same as the attitude she has when she intends to act. But as we have seen an agent can intentionally act despite the fact that her intention has a different content and all this meets a fairly robust consistency requirement on intention and intentional action. Given that it seems impossible to give a satisfactory account of the target example using A3 I propose that we reject A3 altogether in favour of the consistency requirement. If we do this then we allow ourselves to claim that an intentional action does not have to involve an intention to act in that strong way. The theoretical advantage of this move is twofold: first, it gives us a way of understanding practical reasoning that does justice to a wider range of cases that would have previously been considered either irrational or impossible; second, it moves us away from what I consider to be an overly-strong internalist view of the rationality of practical reason (i.e. the PPR view) in favour of a view of practical rationality as self-governance in accordance with norms. I will suggest that moving from one's evaluative judgement to one's intention to act and from here to intentional action is not guaranteed by the sheer logic of practical reason but rather is a regulative, norm-

[13] It is important not to confuse my description of this case with others that can be defended by A3 and the strong belief requirement. The agent is trying *to hit each target*. She is not trying to hit both, trying to hit one out of two, trying to hit one or other or trying just to hit something. In these cases one could argue that she did not intentionally hit T1 but that she intentionally did what she intended to do – i.e. try to hit one or other, both, something etc. I don't want to get side-tracked here in discussing whether someone could plausibly argue that these cases can be handled by A3 and the strong belief requirement. The point is not whether there are cases that meet these requirements but whether there are cases that seem to fall *outside* of these requirements.

governed transition. We will be discussing the normative view of practical rationality in the next section and the implications for the PPR view in Section 4.

So as things stand we have good enough reason to reject A3 in favour of the consistency condition on intention and intentional action. However before we move on there is a clarification I would like to make. When we talk about intentionally B-ing despite intending to A there is a confusion we should avoid. This is not to be read as saying that the agent has the intention to A and then Bs instead. Rather what we are asking when we wonder whether it is ever rational to intentionally B despite intending to A is whether when an agent intentionally acts we can say that she must therefore and necessarily have intended to A. The answer to this question seems to be no, so we know that it is *possible* to intentionally B despite having another intention (I can intentionally go for coffee without intending specifically to do so). Furthermore, with the introduction of the Consistency Condition this does not seem irrational on the part of the agent. Someone might say in response that agents like John, who does not believe he will call Jane, end up being rational with the rejection of A3 and surely this is unsatisfactory and reason to believe that the consistency condition is too weak. But it seems to me that John *is* rational. What is sub-optimal or criticisable about John is that he fails to take cognisance of the fact that he is in danger of violating a norm of rationality that he himself accepts. To continue to doubt whether he will call Jane after he has formed the intention to do so suggests that he is failing to consider the matter closed or he unreasonably continues to doubt his ability to execute his own intentions – both unsatisfactory habits for a rational agent to have. I do not think we can build these sorts of norms into the logic of practical reasoning, but at the same time we do not want to deny that norms and other principles of rational choice guide one's practical reasoning about what to do. What is interesting about the account we are developing

here is that the only rational requirement on intention and intentional action is the Consistency Condition.[14] Although this present account outlines some norms and principles that I believe govern the transitions from evaluation all the way to action this account will be essentially neutral in this regard and makes few claims about the sort of norms that *must* be included. So it remains possible for someone to take this account of practical reasoning and argue that, for example, reflective endorsement is the sole norm which governs all the important transitions such as the one from one's prima facie reasons for action to one's all-things-considered evaluative judgement. All conclusions, according to such a view, must be reflectively endorsed on pain of irrationality *insofar as the agent accepts that norm.* The agent then is irrational with respect to this norm but is not irrational in the sense of logical inconsistency.

However this is not going to satisfy those PPR theorists who prefer to see reflective endorsement as *the* rational requirement on practical reason. One must endorse the conclusion of one's reflection and one must be consistent in one's conclusions about what to do. By coming to a conclusion one endorses that conclusion and it cannot be inconsistent with other conclusions one endorses. But consistency alone only requires that an agent not be formally incoherent. John above is not formally incoherent in his intentions and yet we are not happy to say that he is practically reasoning in an optimal way. The target shooter on the other hand is optimally rational because she meets the consistency condition *and* she successfully makes the norm-governed transition from evaluating the situation to intending and to acting (and all this despite violating A3).

To return to the distinction we raised earlier between rationally and irrationally intentionally B-ing while intending to A, we now can see how such intentional action can be both possible and rational. But in what sense then can such an action also be

[14] It is only by violating this requirement that an agent can be judged formally incoherent or inconsistent.

irrational? If meeting the consistency condition means that agents avoid logical incoherence then there must be more involved in being an optimally rational practical reasoner than consistency. We can say that the consistency condition is a minimal requirement. So we need to say more about the conditions necessary for optimal practical reasoning and as suggested above this can be filled in with an account of the sorts of norms required to ensure the transitions through the stages of practical reasoning (i.e. from consideration of one's reasons to reaching an evaluative judgement and from here to forming an intention and then acting). Weak-willed agents do not suffer from strict logical incoherence or inconsistency because they tend to meet the minimal requirement. But nevertheless they are obviously sub-optimal practical reasoners. So where they go wrong must be in their normative-governance of the transitions involved in practical reasoning. What we need to do now is consider how this normative-governance works in the optimal case.

### 5.3 The Normative 'Gap' and Norm-governed Transitions in Practical Reasoning

Before we move forward let us take a brief look back at the typical transitions that take place in practical reasoning. When we reason about what to do or how we ought to act in any given situation, our practical reasoning involves reflection on our reasons for and against the alternatives open to us. We consider our reasons (or as many of them as we can muster given our own limitations as well as restrictions such as time constraints etc.) and we reach a judgement about how to act. If we fail to reach this type of judgement then we think that our consideration of our reasons is therefore incomplete

and we either continue the process or reach the interim judgement that we should finish this evaluation later. This much seems hard to deny – there is something decidedly peculiar about an agent who reflects upon her reasons for action and just picks one as the reason on which she must act. We do not typically assume, then, that we can decide how to act by choosing any old prima facie reason. Let us say then that when an agent has considered all her reasons for and against a course of action and has reached a judgement about how she should act she has an *evaluative judgement* about how to act.

On the face of it it seems natural to assume that an agent's intention to act will have the same content as her evaluative judgement. After all, if Jane considers all the reasons and judges that the best thing to do is to spend the day writing, surely then that is what she will *intend* to do? However I do not believe we can assume that this is always the case, and this is where our weak-willed cases are important. How are we to explain cases where the agent reaches an evaluative judgement and yet forms the intention to do something else – something that she has already considered and overruled? It could be said in response to this that reaching an evaluative judgement *just is* forming an intention to act and the agent then acts intentionally. Given that we have argued that one can intentionally B despite intending to A it could further be claimed that weakness of will can be explained as *irrationally B*-ing when one intends to A. But this does not seem to provide an accurate explanation for all cases. We are familiar enough with situations where an agent might believe, say, that exercising now would be best but does not intend to exercise now. She might intend to stay in bed; that is to say, she forms an intention based not on her evaluative judgement about what is best but on a reason she has already considered and overruled by the reason she eventually adopts as her evaluative judgement. This case is as much an example of weakness of will as is irrationally B-ing despite intending to A. The difference

between the two cases is that in one the agent forms an intention and then intentionally does something else, and in the other she does not form the *right* intention (i.e. the intention that her evaluative judgement suggests she should form).

The best way to account for these cases is to distinguish the stages of practical reasoning and suggest that transitioning from one stage to the next is not logically guaranteed. Let us assume that agents reach evaluative judgements in the same manner they reach conclusions in ordinary theoretical arguments based on premises they accept. We reach new beliefs in this manner and just as beliefs are subject to consistency and coherence requirements so to are our evaluative judgements about how best to act. The transitions we are particularly interested in are the move to the intention to act and from there to the intentional action. If an agent does not make these transitions from the sheer force of logic then how does she make them? Let us first consider the move from evaluative judgement to intention.

Both the ordinary rational case and the weak-willed case give us reason to believe that the content of the evaluative judgement and of the intention do not have to be identical. If evaluative judgement is always of the form 'X is the best thing to do all-things-considered' (and we really have no reason to think otherwise) then rational agents, on certain views, would always act on intentions of the form 'X is best, do X'. But clearly this is not matched by our experience of acting on an intention. An agent can evaluate concerning which option is better – the devil or the deep blue sea. She might decide that the deep blue sea is the best option, all-things-considered but when she forms her intention she will not have the intention 'deep blue sea is best, go to the deep blue sea' , she will just have the intention 'go to the deep blue sea'.[15] The weak-

[15] It just does not seem plausible for an agent in such a situation to say that the option she chooses is best. Of course there is the trivial sense in which this is true – insofar as one always has to choose it is better to choose than not to do so, and so in this sense the 'deep blue sea' option will be the best of a bad lot. But this is not the same as believing one's choice to be best. Sophie must choose which of her two children must die. She chooses child 1 but she does not think it *best* that child 1 die.

willed agent on the other hand might well have the evaluative judgement that 'X is best all-things-considered' and yet do Y without believing it to be best. Consider Austin's dessert example. The agent judges that given her reasons it would be better not to have a second helping of dessert but she goes ahead and does so anyway. Now if we want still to claim that an agent's intention is always of the form 'X is best, do X' then to be consistent we have to say either that the agent must really have judged that two helpings was best all along (and so her evaluative judgement and her intention have the same content) or that she had the evaluative judgement that 'one helping is best all-things-considered' and yet ended up acting on the intention that 'two is best, have two'. The first option does not allow for weakness of will (as described) at all and the second portrays the weak-willed agent as involving herself in terrible logical error.

Neither of these options is attractive and so it might be best to re-examine the connection between the content of one's evaluative judgement (all-things-considered) and one's intention to act. However, rational agents do typically intend to act according to their judgement of how best to act, so how does the agent transition from one to the other? To explain this I shall borrow two notions – Davidson's idea that our practical reasoning operates on a *principle of continence* and Bratman's idea that a *commitment* is involved in coming to have an intention to act.

### 5.4 Phronesis and Commitment

Michael Bratman makes a crucial point about intentions. He argues that intentions are not just like more reasons to be considered in an agent's practical reasoning. Rather, intentions I have constrain my practical reasoning about what to do because they determine what reasons are relevant in my planning and thinking about what to do. For example, if I intend to fly to Miami next Tuesday then any reasons I may have in favour of visiting friends in Australia on Tuesday will be ruled out of my practical reasoning about what to do since they are *at least* not compatible with, or not relevant to, my confirmed intention.[16] Contrast this view with the PPR criticism of the IR view we met earlier – that the IR view allows any agent to be are rational since IR theorists believe that agents always act as they intend and so any agent with an intention must be said to rationally hold that intention.

If we stop identifying evaluative judgement with the intention to act then how are we to explain the transition from the former to the latter? Since weakness of will requires us not to identify the two the most plausible suggestion is that this is a norm-governed transition. But what sort of norm plays this governing role? Rational or optimal practical reasoners tend to intend in accordance with their evaluative judgement as to how to act. When everything runs smoothly, we find ourselves evaluating our reasons for action on the assumption that we will act accordingly. It is difficult to imagine a proper-functioning agent who assumes that she will not act in accordance with her considered judgement about how to act – the only circumstance in which this might happen is when an agent knows that she *cannot* act according to her judgement (perhaps because she is prevented from doing so by forces outside her

[16] Note however that I do not have *an intention* to visit friends in Australia. This would make for a case of incompatible intentions and in this situation the one intention cannot determine the admissibility of the other since they are equally intentions that I hold about how to act on Tuesday. In such a case the agent needs some further intention about her future actions that can resolve the difficulty – and this case is different again from that of the agent trying to form an intention to act upon consideration of equal but opposing reasons for acting.

control). But even in such a case the agent's knowing this means that she will include this information in her evaluation of her situation, and this will not be a case of her assuming that she will not act in accordance with her judgement *tout court*. Let us take it then that such an agent would be a sub-optimal practical reasoner and that agents more generally evaluate their reasons on the assumption that they will act accordingly.

However we are not envisaging any *mere* assumption here. This assumption plays a guiding role in practical reasoning – the agent moves from thinking that *this* is how she ought to act (all-things-considered) to intending so to act. Presuming the agent then goes on to perform the act in question[17] we should be quite happy to say that we have an agent with practical wisdom, or a *phronimos* as Aristotle would call her. This assumption that guides practical reasoning is more appropriately called a *norm* of practical reason since it is a principle that the agent accepts as guiding or governing her intention-formation. To say that this is just an assumption fails to capture the agent's acceptance of and *commitment to* this principle whereas a *norm* is something an agent in some sense adopts as her own when she conforms her practical reasoning to it. An assumption could get an agent's commitment and acceptance but it is equally true that an assumption is something we could just 'try out' in practical reasoning without accepting it as one's own guiding principle. To acknowledge this difference let us call this assumption here the *Phronetic Norm.*[18] An agent accepts the phronetic norm as guiding her practical reasoning only if she forms an intention to act in accordance with her all-things-considered evaluative judgement about how to act.[19]

---

[17] This is not a presumption I want to make in general since I think that there is further work to be done in explaining the move from intention to intentional action. But since we will be coming to this shortly we can just go with the presumption for now.

[18] Davidson discusses what he calls a "principle of continence" (act in accordance with your best judgement all-things-considered) and argues that this principle governs practical reason despite not being strictly rationally required. Similarly Bratman discusses a principle of practical reason "PR" which has a similar function although it does not require that the agent act in accordance with what she thinks is strictly *best*. See Davidson (1980) p.41; Bratman (1979), p.165.

[19] Strictly speaking when an agent reaches an evaluative judgement she only has to *intend to act* accordingly if she is thinking about how to *act*. We are dealing with *practical* reason and so in the case

The next issue we have to consider, however, is how we are to understand an agent's acceptance of a norm, specifically a norm governing intention-formation. We know that if an agent intends in accordance with her evaluative judgement then she is guided by the phronetic norm – but how does this work? I think we can most easily see this via the agent's dispositions. The agent who is guided by this norm makes a *commitment* to intend according to her all-things-considered evaluative judgement is reflected in her disposition *so to act*. For example, an agent's commitment to A-ing in circumstance C is reflected in her disposition to A when in C, and she will criticise herself when (all things being equal) she fails to A in C. The normative gap that is potentially present after reaching an evaluative judgement is clearer with an example. Imagine two agents, Bill and Bob who each judge that they should not take the camera even though they could easily get away with it. We can also imagine that as the situation unfolds Bill intends in accordance with his judgement but Bob does not and he takes the camera. Despite having the same judgement (in the same circumstances) only Bill intends as he should if he were guided by the phronetic norm. But we do not want to say here that any agent with a commitment to act must therefore have an evaluative judgement to the effect that that action is the one to be performed *all-things-considered*. An intention to act really is a stable commitment to act but since weak-willed agents can have stable intentions that do not match their evaluative judgement we do not want to say that a commitment always matches the evaluative judgement.

A similar example shows this gap from the other direction: Bill and Bob each have the intention to stay in bed today but this does not mean that they both have an

---

of action the agent does need to intend in accordance with her judgement; but an agent might be thinking about something else, such as how to react to some event rather than how to act, and in this case she might develop as a response a certain attitude that is in accordance with her evaluative judgement. This transition from evaluative judgement regarding X to attitude to X is governed by a similar norm concerning forming an attitude or response in accordance with one's evaluative judgement. Given the similarity, and for simplicity's sake, let us consider the phronetic norm as covering such cases of non-action also.

evaluative judgement favouring staying in bed. Bill has such a judgement and so his intention is formed in accordance with the phronetic norm, yet Bob has no such judgement. He has an intention to act that does not match the evaluative judgement he has (for example the judgement that he should get up and go to work). What we can say of Bob is that he might be weak-willed. But why only that he *might* be; why are we not so sure that he definitely is weak-willed? The reason is that from the scant information given we cannot tell whether he is violating the phronetic norm or not. If we expand on Bob's reasoning I think we can see things more clearly. Bob has spent the last number of days trying to decide how to act this particular morning – should he get up early or stay in bed? He has no prior intention either way and so his evaluation of his situation does not include any reason such as "I'm committed to getting up early in circumstance X and this is circumstance X". Bob concludes that he should get up early but he ends up with an intention to stay in bed. Given Bob's situation I'm reluctant to argue that he is simply violating the phronetic norm despite the appearances. It is true that he is not intending in accordance with his evaluative judgement but in this case his judgement does not really settle what to do. Since he has to act one way or the other then he must decide on an option, but since he has no reasons in favour of one that override his reasons in favour of the other he is in a situation of arbitrarily choosing for the sake of choosing. He ends up making a judgement that is not grounded in real commitments he has (beyond a general commitment to action itself) and so he judges that he should get up early but then intends to stay on in bed. Imagine another scenario – one where Bob has prior intentions regarding getting up early in circumstances such as these. These intentions are among his evaluable reasons and so by his own lights he really should intend to get up early. In this situation we can say that Bob is weak-willed[20] because he has reasons

[20] Bratman makes a similar point about the underdetermination of commitment by value judgements.

for action that he has already committed himself to in earlier practical reasoning and so to commit himself to some course of action that is overruled by these standing commitments is to have an intention to act that is not supported by his reasons for action.

It seems then that to determine an agent's practical rationality we need to examine the extent to which she intends to act in accordance with her judgement about how to act where that judgement is the result of consideration of all her evaluable reasons and where these include her relevant standing commitments. And in the rather unusual instances (such as Bob's above) where an agent might not have any relevant standing commitments she cannot be called weak-willed for reaching an intention to act that does not match her rather arbitrary judgement about how to act. I do not think this case is a very typical case at all since most of us when making decisions about action have some kind of relevant standing commitments that get considered in the evaluation of our reasons. However let us nail our colours to the mast here – are agents in this (albeit unusual) situation guilty of practical irrationality? It seems to be a rather harsh accusation, especially since the agent sees no real conflict involved. Nevertheless it is hardly a disposition we would expect a rational agent to have. It does not quite amount to impetuousness but it seems more akin to it than to a stable disposition to follow the phronetic norm. But one might wonder why we are talking about stable dispositions here at all. An agent makes the transition from her evaluative judgement to a corresponding intention to act and we say that she is so *committed* to action. In cases where the agent fails to so intend simply because has no relevant standing commitments regarding the judgement she eventually reaches why not say that there just is nothing at stake for the agent and her intention does not have to be determined by her evaluative judgement?

---

However he argues for a tighter connection between evaluative judgement and intention.

The problem is that this picture plays down something extremely crucial – the agent's self-direction or governance. An agent reaching an intention about how to act is making a decision or a commitment to act in that way and the phronetic norm ensures that the move from consideration of reasons to the intention to act is a piece of practical reasoning. The agent with a decision or commitment regarding action or regarding how she should make decisions at a later date is involved in *governing* herself through making commitments. As we have seen reaching a judgement does not guarantee that the agent will form an intention accordingly – it is the phronetic norm that ensures this.

An agent who is involved in directing her actions is going about the business of governing herself. When agents evaluate their reasons for action (which include their own desires) they evaluate them with reference to their own standing commitments. The process of coming to have a standing commitment just is this same process of coming to have an intention not just about how to act but for example about how to regard reasons for action one may have in the future or whether to approve of certain events, etc. An intention such as this plays the role of a standing commitment when the agent guides her practical reasoning in accordance with it. A standing commitment is formed when the agent is deciding not just how to act *right now* but how to regard certain kinds of reasons (including desires) in her practical reasoning.[21] She reaches such a standing commitment when she does not have any more doubts about her evaluative judgement, so sudden change suggests that she has not evaluated properly or to her own satisfaction. An agent who changes intention even though strictly speaking her intention remains underdetermined by her evaluative judgement is not weak-willed and we cannot say she is irrational, but we can have doubts about the stability of her

[21] cp. Harry Frankfurt 'Freedom of the Will and the Concept of a Person' *Journal of Philosophy* LXVIII, 1 (1974), also *Necessity, Volition and Love* (New York: Cambridge, 1999).

intentions and standing commitments. We will come back to this issue below when we take a look at *stability* of intention.

Let us take a look at a situation where the evaluation an agent undertakes is about her desires, that is, the agent evaluates them because she wants to know whether she should act on these desires generally as a standing commitment or self-governing practical policy. By forming and following a standing commitment about these particular desires the agent willingly places constraints on her further intention formation. For example, Jane desires to go to the concert on Saturday but she worries that it will cost more money than she should responsibly spend. Her desire to go is at least a *prima facie* reason in favour of going but her fears about the cost are equally *prima facie* reasons against. If Jane has no overriding reasons in favour of one option or the other then she is in a situation where the only pressure on her is the pressure to decide how to act. It is really not possible to judge the agent's capacity for self-governance on the basis of these sorts of decisions.[22] But what about the more typical case in which Jane would indeed have standing commitments that can act as standing commitments in her evaluative reasoning? Let us suppose that Jane decided last year to be cautious and prudent with her money and as a practical rule she decided that when in doubt about an expense it is best not to spend. In her practical reasoning about how to act this rule has constrained Jane's evaluations – because she has this rule she has determined that "...but I really want to X" is not an overriding reason in deliberations about expenditure. We can say that she has a standing commitment concerning cautious spending and with a clear commitment like this she can easily override prima facie reasons of the "but I want to" variety. However Jane might also have standing

[22] Of course Jane might decide to go to the concert (evaluative judgement) and then change her mind (reach a different intention), like the agent Bob above. In this case she might not be weak-willed because her intention is not in accordance with an evaluative judgement that takes standing commitments into consideration. But we still might have reason to question her capacity to sustain intentions or commitments she does form and thus her general capacity for self-governance is in question also.

commitments concerning the value of music and concert-going. Perhaps she thinks that attending concerts adds cultural richness to her life. This standing commitment might be in conflict with the cautious spending commitment, leaving Jane with a more difficult evaluation. Let us say she judges that she should go to the concert and reaches the corresponding intention to act. In this case Jane intends to go to the concert because she is satisfied that the evaluative judgement she has in favour of doing so results from the overriding standing commitment she has favouring acting to pursue cultural experience. Jane treats her attitude in favour of the value of cultural experience as her justifying or overriding reason in motivationally efficacious practical reasoning. If Jane was weak-willed she would not intend or commit herself to acting in accordance with this justifying reason. In that case the relation of her standing commitments remains the same and so Jane will reach the same evaluative judgement, but that judgement will not be motivationally efficacious – Jane will not intend accordingly. Despite satisfying herself about which reason she should act upon Jane intends to stay at home and not spend her money. She is in clear violation of the phronetic norm here (i.e. she fails to intend in accordance with her evaluative judgement) and it is this which constitutes her weakness of will – her intention is in direct conflict with the standing commitment she considers overriding. If it was in conflict with her evaluative judgement only then we would not be able to tell whether there is weakness of will without figuring out whether she had standing commitments regarding her reasons (she might be in the arbitrary decision situation). So the weak-willed agent fails to intend in accordance with her evaluative judgement – judgement that is the outcome of allowing her overriding standing commitment guide her practical reasoning. Jane is entirely culpable for not letting that evaluative judgement be motivationally efficacious. When Jane is not being weak-willed she is effectively engaging in practical reasoning that

leads her to reject a certain reason (in this case a desire) she has. By intending in accordance with her evaluative judgement she endorses both the intention itself and her standing commitment in favour of it.

By having standing commitments an agent gives herself a structure of self-governing policies that will constrain any further practical reasoning. If some consideration is to be even a *prima facie* reason for action then it must not be inconsistent with or irrelevant to the agent's standing commitments. To this extent agents are not crude instrumental reasoners who reason about the relevant means to a desired end. Agents can distinguish between ends that are only justified by prima facie reasons and ends that are consistent with existing standing commitments and intentions. Agents can also engage in practical reasoning about her commitments and intentions and these are not easily shifted by fleeting desires the agent may have. In fact what the agent believes and desires is often determined by her commitments and intentions. If Jane has a standing commitment regarding the value of cultural experience then she will be more likely to desire to attend cultural events and to believe that these are worthwhile. Her standing commitments function as stable and clearly articulable attitudes which determine what is an admissible prima facie reason in practical reasoning about how to act.

So far then we have suggested that an agent's transition from her evaluative judgement to her intention to act is practical *reasoning* because it is explained by a principle of such reasoning we have called the *phronetic norm*. This norm guides practical reasoning but does not determine the content of the agent's reasoning. What plays that role are the standing commitments the agent has that favour or reject certain actions over others. A standing commitment is a clear attitude regarding the value to the agent of a certain course of action and it is reflected in the agent's disposition to act

in accordance with her evaluative judgement.

### 5.5 Sustaining commitment: Stability

If an agent is to be a reliable and self-governing practical reasoner who can sustain her standing commitments and intentions in the face of such things as incompatible desires then it is clear that *stability* must be a feature of these commitments and intentions. An agent such as Jane in the example above has a standing commitment to favour cultural events and she holds this commitment over time and through periods of changing life circumstances. A standing commitment to vegetarianism is another common example. Vegetarians tend to be steadfastly committed to abstaining from meat and meat products and they carry this as an important standing commitment in all their practical reasoning about diet, food purchasing and preparation and so forth. An agent who claims to be a committed vegetarian who then is found to be eating meat when there is a vegetarian alternative is someone whose claim to her standing commitment we would probably question. But the attitude that comprises an agent's commitment is an attitude about the value of *X* in the agent's life – it does not have to include an attitude about the value of stability. That is, a committed vegetarian is committed to abstaining from meat except in very specific (probably life-threatening) circumstances but she does not also say to herself "...and I'm committed to having stable commitments". Stability is a feature of her commitments, not a necessary further commitment in its own right that must *accompany* all other commitments.

An agent with standing commitments then will have those commitments until

sufficient new information forces her to re-evaluate. But this new information must be of a significant magnitude; for example for some vegetarians a discovery that animals do not suffer at all in intensive feedlots might be a new consideration of sufficient magnitude to cause an alteration of standing commitment. But so long as an agent has a commitment her practical reasoning will be constrained by it. However if an agent is to have lasting and sustained commitments that determine the admissibility and relevance of her prima facie reasons then she will need to have certain dispositions towards her commitments that determine their stability. But what sorts of dispositions?

Most importantly, an agent must have the disposition to practically reason in accordance with her standing commitments. An agent forms a standing commitment when she reaches an intention concerning the values she wants to constrain her practical reason in the future. That intention will persist in the agent's practical reasoning so long as she does not care to revise it (because she has no reason to do so) and so long as she regards that commitment as *self-governing.*[23] There is a view that when an agent has stable standing commitments she intends to be motivated accordingly.[24] However I think adding in a further intention here might lead to a complicated regress of intentions. If I must intend to be motivated by my commitments then I must treat that intention as a necessary reason or consideration in my practical reasoning. Then I have a commitment, an intention regarding my commitment and what is to stop the regress that would see me have a further intention regarding how my intention to be motivated by my commitment is to work etc.? Practical reasoning that is constrained by standing commitments does not seem to require the agent to have reasons that are *about* intentions regarding those commitments. Instead agents seem to practically reason in accordance with their commitments and they do this because they

[23] In this I am influenced by Michael Bratman's comments on self-governing policies. See Bratman (2000) pp.51-5.
[24] For example, see Tiberius (2000), p.431.

have dispositions to retain their commitments, dispositions to constrain new intentions in light of their commitment and generally have the disposition to reason in a way that is constrained by their standing commitments.

What we are claiming then is that to regard a commitment as self-governing is to have a pre-disposition to treat the evaluative judgement that results from practical reasoning involving such commitments as motivationally efficacious. This is not to say that agents do not feel the pull of contrary desires or see the attractiveness of other and incompatible reasons for action. Agents do feel these other influences on their practical reason but because they have standing commitments and predispositions to be motivated or guided by them the other influences usually lose out. It is important to see the role that commitments play in practical reason. If commitments were simply the agent's current desires or some such state then (as Korsgaard rightly points out) they could not constrain the agent's practical reasoning as they do. Instead by having standing commitments an agent can come to reject certain desires she may have – desires that are incompatible with practical reasoning in accordance with her standing commitments. If commitments were just current desires, for example, the agent would have no reason to reject a later desire on the basis of them, and in this case no agent could ever act against her judgement of what to do. We have seen such criticisms already and expressed concern over them. But this account of practical reason sees standing commitments as having a role to play in their own right.

Agents gain a conception of their practical identities by identifying with their standing commitments. Here again we are in agreement with proponents of the PPR view who make similar claims. However I do not want to go further and argue that agents must endorse some general principle *as* a conception of their identity. The agent sees herself in terms of her standing commitments and this develops over time as more

and more decisions about action are reached in accordance with those commitments. The vegetarian sees herself more and more as a committed vegetarian as she practically reasons over time in accordance with her commitment. She has a disposition to reason and be motivated accordingly and so what might have begun as a tentative commitment to certain dietary restrictions gradually becomes a defining characteristic of the agent herself (qua agent). We might even say that the stability of the commitment increases over time also, as the agent consistently faces situations where she is satisfied that reconsideration of her commitment is not required. So the agent who says "I am a vegetarian" in response to a question about her crucial standing commitments has a stable commitment to the relevant principle, intends to act accordingly and rejects desires or other prima facie reasons she has that are inconsistent with her commitment.

But having commitments does not guarantee that an agent will form the *intention* to act accordingly. What this does guarantee is that an agent will reach the corresponding evaluative judgement. We can easily imagine an agent, Jane, who has a stable standing commitment regarding being charitable where one can. When she evaluates whether to give to charity today since she has some surplus money in her pocket she will conclude that she should do so. As a self-governing agent with the disposition to constrain her reasoning in light of her standing commitment Jane evaluatively judges that she should give the money away. Her self-governance also includes the disposition to be motivated by her evaluative judgement so all things being equal we would expect Jane to intend to give the money away also. Jane has this latter disposition because her commitments are generally formed and regulated in accordance with the phronetic norm. Finding one's evaluative judgement to be motivationally efficacious reflects one's commitment to the phronetic norm in the process of intention-formation. The evaluative judgement is authoritative for the agent because it is the

result of practical reasoning from her standing commitments and relevant prima facie reasons, that is, it is consistent with her commitments.

By taking her evaluative judgement to be motivationally efficacious Jane shows that she has a further attitude favouring intending in accordance with her evaluative judgement. This means that the phronetic norm is another standing commitment of the agent. Now it might be objected here that this seems to make this norm very much an optional matter for the individual agent and as such this is far too weak an account of practical reasoning, since forming the correct intention should not depend upon whether one is committed to doing so or not. But as we have seen weakness of will and cases of underdetermination of intention complicate matters for those who would argue that the transition from evaluative judgement to intention is direct. If we think it is possible that the intention to act is strongly entailed by the evaluative judgement then we would have to conclude that the content of each must be the same in all cases. But this makes weakness of will seem either completely impossible or an error of extreme illogic on the part of the agent. However cases of weakness of will have shown us that the content of the evaluative judgement and that of the intention to act do not always match each other[25] and so if we are to see the transition from the one to the other as practical *reasoning* at all then we have to assume that what we have here is a norm-governed transition. What we need to keep in mind is that while it is possible that an agent might reject the phronetic norm this does not mean that she is beyond our criticism. What makes an agent a rational *practical* reasoner is her acceptance of the phronetic norm, which thus functions like a principle of rational choice. So the agent who violates the phronetic norm can be accused of being irrational (in the sense of being unwise) but not of being logically inconsistent. It does not seem possible for any agent who considers

[25] An agent can have an evaluative judgement regarding A-ing but for whatever reason not have the intention or commitment to A; and it's also possible for an agent to have an intention that does not have the same content as her evaluative judgement.

herself to be self-governing to reject the phronetic norm since to do so means that the agent is making it impossible for herself to live up to the basic features of practical reasoning: namely, that practical reasoning involves coming to a conclusion about how to act based on consideration of her reasons, and that the process of consideration is not complete until the agent figures out what she finds to be her best or most satisfactory reason upon which to act. All the phronetic norm really does is articulate the agent's recognition of these basic features of self-governance. So if I recognise that these are the basic features of practical reasoning and I proceed to reason accordingly then I am allowing these basic features be authoritative or normative for me. An agent who wants to reject this can hardly be called a practical reasoner at all for it is hard to imagine what it would be like to practically reason without going through the process of gathering reasons for and against, examining them carefully and finally coming to a conclusion about what to do all-things-considered. We could alternatively go on to say that such an agent would not be a rational practical reasoner but I'm not sure what extra work the word 'rational' really does for us here.

There is another issue regarding stability we should discuss and it concerns the weak-willed agent and the stability of intentions. Despite reaching an intention to act that does not correspond to her evaluative judgement, the weak-willed agent can go on to act on that weak-willed intention. Now to do this she requires her intention to be *stable* in its own right. In order to make the transition from intention to intentional action the agent needs to sustain the weak-willed intention long enough to see it through all the way. This underlies the importance of the phronetic norm. If an akratic can come up with an intention based on a lesser reason she has and that intention is stable then stability is not the feature of intentions that determines their rationality.[26]

[26] It is not a sufficient condition on intention but it is a necessary one. On the other hand, following the phronetic norm is necessary but not sufficient; an agent who makes commitments that are in line with her evaluative judgement but are not stable commitments is weak-willed because she will not hold that

Stability is a crucial feature nonetheless. Stable commitments and intentions allow agents to make plans regarding future action-decisions and to coordinate their lives over time. Agents live and choose in accordance with their commitments – they become part of the practical identity of the agent. So to put one's commitments constantly up for revision brings instability and in turn difficulty coordinating the direction and choices in one's life.[27] Weak-willed agents can have stable intentions and can sustain their weak-willed intentions. They can get into situation where they make a commitment to a lesser reason for action and their commitment to this reason is reflected in their dispositions to be motivated accordingly. For example, as a lesser reason an agent may have 'it's OK to allow oneself to be distracted by short-term pleasures'. This is overridden by her evaluative judgement which, after consideration of all her reasons might be something like 'work towards my long-term goals without distraction'. She ends up not intending in accordance with this (and so violates the phronetic norm) and forms an intention to take advantage of all the coming pleasures. This intention can be reliably stable – this stability is what enables the agent to move from *intending* to *intentionally acting*. Without stable intentions an agent would function sub-optimally. But it has to be noted that an agent can form a stable intention even though she has violated the phronetic norm and that she can form an intention that *is* in accordance with the phronetic norm but it may lack the stability required to see it through to action. So acting in accordance with one's evaluative judgement I'm calling a *norm* of practical reason; but stability is a necessary feature of commitments that is reflected in the *disposition* of the agent not to hastily revise her commitments to action.

---

commitment long enough for it to be motivationally efficacious. The agent who forms a stable commitment where that commitment is not based on the content of her evaluative judgement is also weak-willed since she will go on to act on a commitment that she knows on reflection that she should not have formed. An agent is weak-willed then, if she violates *either* of these conditions.

[27] Bratman has discussed the issues of coordination and planning at length in (1987) chapter 3 and 4.

### 5.6 Weakness of Will and the Authority of Practical Reason

In what sense does weakness of will constrain an account of practical reason? What about those PPR criticisms? We now have the resources to give a better explanation of cases of weakness of will. What we have seen is that there are situations where agents make judgements based on their valuations, desires, beliefs (or general reasons for action as we have been calling them here), but coming to such a judgement does not guarantee that the agent will form the corresponding intention and be motivated to act. By examining cases of weakness of will we have seen not that these agents are exceptions to the rule but that the 'rule' has not been fully adequately conceived in the accounts previously examined. What further bolsters our case is that cases of ordinary, non-weak-willed practical reasoning can also lead to similar problems – two agents can reach the same judgement but only one might go on to form the relevant intention and be motivated to act.

As a view about the agent's authority over herself and her own actions we can quite satisfactorily argue that a rational agent who forms a stable intention to act on her evaluative judgement has a commitment to take her all-things-considered reason for action to be motivationally efficacious. The weak-willed agent will do one of two things: she might make a commitment that is in line with her evaluative judgement but is not a stable commitment so she is weak-willed because she will not hold that commitment long enough for it to be motivationally efficacious. Or, she alternatively she might form a stable commitment where that commitment is not based on the content of her evaluative judgement. This agent is weak-willed since she will go on to act on a commitment that she knows on reflection that she should not have formed.

But perhaps the PPR view we examined in previous chapter might still have doubts about this account. The charges of descriptivism and of permitting the agent to

always be instrumentally rational do not apply to the view we have been sketching here. We are not merely describing how agents *are*; rather we are trying to give an account of how agents authorise their own actions. It is obvious also that we are not saying that any agent is rational insofar as she ends up doing what she in some sense *wants* to do. It is clear that we are saying that agents who do not form stable intentions violate the phronetic norm, do not form a commitment regarding how to act that is guiding and efficacious and they often lack the stability necessary to form and hold motivationally efficacious intentions. But the PPR view might still wonder whether we have really captured what it is to *endorse* a reason for action as one on which the agent must act on pain of irrationality.

It might be objected that we would be better to think of standing commitments as special *beliefs* about the value of certain courses of action rather than as *attitudes* that are reflected in the agent's dispositions to act. What might provoke an objection of this sort is the idea that attitudes cannot by themselves provide the sort of normative guidance agents need in order to figure out how they ought to act in any given situation. Beliefs and reflection on them about what an agent has most reason to do can be relied on to guide an agent to the most appropriate action. What lies behind this idea is the Kantian idea that to figure out which of my desires to endorse as a reason for action I have to stand back from my desires and attitudes in order to do the necessary evaluation of them (i.e. in order for me to reach a belief or judgement about what I have most reason to do). However by arguing here that we need to add an (attitudinal) account of commitment to any account of practical reasoning I do not want to give the impression that agents form authoritative or guiding commitments or intentions to act that are predominantly determined by whatever desire or attitude they happen to have at the time.

What I am suggesting is that we accept that evaluative judgement sometimes does underdetermine both the agent's commitment or intention to act and her intentional action. This is because what constitutes commitment to action is a favourable attitude towards the content of the evaluative judgement. This attitude is reflected in the agent's disposition to act in accordance with her commitment. Stability and persistence are key features of these commitments. I also suggest that rational agents have a commitment to intend in accordance with their evaluative judgement. We see this commitment reflected in the rational agent's disposition to make a commitment to act in accordance with their evaluative judgement (with the exception of certain circumstances where it would not be irrational to fail to do this – i.e. in cases where there is no reason to favour one option over another). This is a commitment it makes sense for agents to have and so I have called this the phronetic norm. A stable and persisting favourable attitude towards the contents of one's evaluative judgement can hardly be considered a mere fleeting desire. Our commitments are thus not like a simple desire for X.

There is a major role for deliberation and evaluation in the formation of our commitments. We can come to have new stable commitments as a result of our evaluations that in turn will go on to influence further practical reasoning. These commitments also act as standards that determine the admissibility of other desires and attitudes in our practical reasoning. The vegetarian agent we discussed is a case in point here. Her commitment to vegetarianism is a stable and persistent commitment and having this means that she does not even consider desires she might have (for example, for a pair of leather shoes) as offering her justifiable reasons for action. If she begins to have doubts about the stability of her commitment to vegetarianism then perhaps she is beginning to reconsider and reflect upon the commitment itself.

In making these comments I have not argued that we *cannot* see commitments as beliefs rather than attitudes. Instead what I hope is clear is that an attitudinal account of commitment can overcome the objections placed before it. However I think we can go further and argue that we can also understand having an evaluative judgement in terms of having certain kinds of attitudes. But this is another entire topic in its own right and as such I cannot do justice to these issues here. What I have done instead is include in this account of practical reasoning evaluative practical reasoning in a form that PPR theorists could approve of. Agents reflect on their reasons for action and then reach evaluative judgements. I have accepted this instead of accounting for evaluative judgement itself in terms of attitudes. What I have disputed here is their view that endorsement amounts to evaluative judgement. Here I suggested that attitudes *do* play a crucial role in endorsement. I do not believe that beliefs could play this role. To be committed to some course of action one needs more than the belief that this is the best or right course. Commitment or intention to act can be present even when an agent does not believe, or has real doubts about her action. It is by examining cases of weakness of will that we come to see that reaching an evaluative judgement (that is, a conclusion of one's practical reasoning) is not the whole story when it comes to the endorsement of reasons. Even if someone continues to argue that no kind of IR view (and not the even 'Reflective Dispositionalism' view presented here) can work as a theory of practical reason they still face a difficult obstacle when it comes to weakness of will. One does not need to accept an IR view in order to see that there are problems with the assumption that reaching an evaluative judgement is constitutive of endorsing a reason for action. This is not 'motivational skepticism' as Korsgaard puts it. What we might have to accept however is that there is a transition in practical reasoning that must be made between an evaluative judgement or conclusion of reflection and the

commitment or intention to act. But this transition is not one that the agent is *rationally bound* to make. The agent who fails to make the transition is open to criticism however. We can say that she is unwise or that she violates the basic conditions of practical reasoning. What she is not doing however, is violating a law of reason binding on all rational agents as such.

## Conclusion

We began our examination of the problem of weakness of will by recognising two undeniable intuitions about action. We believe that (very broadly speaking) actions follow upon some sort of judgement or consideration and, secondly, we also believe that weakness of will is possible, that even when there are no obvious impediments, agents can and do knowingly fail to act as they think they should. On the face of it, these seem difficult to reconcile. If we act according to *reasons* then it seems problematic to say that *ceteris paribus*, we can just fail to act accordingly. At this point the options appear stark. Either we argue that the idea that reasons are directly motivating is too strong or we reject weakness of will as traditionally understood. Neither option seems likely to be satisfactory. If we hold the view that reasons are not directly motivating, that motivation comes from a further 'push' from other contingent states of the agent, then it seems that at least agents cannot be immediately considered weak-willed unless they fail to act despite having the reason *and* the relevant independently intelligible and contingent motivation. But this does not get us beyond the problem of our apparently inconsistent intuitions. There is no solution to be gained by arguing that reasons are not directly motivating, for either way the same difficulty remains. If we hold the view that reasons are directly motivating then weakness of will seems impossible – lacking the motivation to act is more plausibly thought to be due to not having the reason to act in the first place. Yet a similar problem arises if we say that reasons are not directly motivating, for in that case it is more plausible to say that failing to act is due to lacking the independent motivator. So it seems that if we argue either that reasons are directly motivating or that they are not we still have the same problem: weakness of will as traditionally understood seems impossible.

One obvious solution to this problem is to reject weakness of will, or at least, reject the notion that it is possible to *knowingly* fail to act as one thinks one should or considers best all-things-considered. One could argue, as we have seen, that what agents really suffer from is some form of self-deception or some psychological failure. Perhaps the agent who fails, for example, to study for an important exam is deceived as to the real weight she has given her reasons supporting studying. Perhaps she suffers from apathy or depression, or maybe she is cynical due to feelings of hopelessness or low self-esteem. This might not be an unpalatable response if one doubts altogether the capacity of a theory of practical reason to explain how our actions or deliberation about actions could be directed by reasoning about what to do. However this is not an issue we have set out to explore in this work. If we want to pursue a theory of practical reason then weakness of will poses a real problem. If the view is that reasons are motivating then cases of weakness of will as traditionally understood appear to be counterexamples. If we redescribe weakness of will to avoid this difficulty then we run the risk of failing to save the phenomenon. Again this might not universally be considered an unsatisfactory response to the problem. We might say, as we have seen some Kantians do, that agents are motivated by their reasons only in so far as they are *rational* agents. Thus, no rational agent by definition would ever knowingly fail to act against her all-things-considered reasons for action. Any agent that does is practically irrational and we can call *this* weakness of will if we so wish.

Yet as we have seen, it has not been satisfactorily shown that a rational agent who has a reason to act all-things-considered is also motivated to act accordingly since we can point to underdetermined cases concerning rational agents whose actions do not seem to be the direct result of reaching a judgement about how to act. Consequently I have argued that following the correct reasoning procedure for figuring out how best to act (as suggested by the PPR view) cannot be guaranteed to bring a rational agent to

act. What weak-willed actions have in common with these underdetermined cases is that in both cases the agent does indeed seem to reach a conclusion based on their deliberation or reflection on what to do and yet their action does not seem determined, or at least not entirely determined, by this conclusion.

Since these agents reach an evaluative conclusion based on their considerations it is implausible to deny that these agents are practically reasoning correctly and so it makes no sense to deny that they have reached a proper conclusion after consideration of their reasons. We *can* say that they have genuine judgement because they have evaluated and reached a conclusion based on their evaluation of their available reasons. So far there is no difference between the ordinary rational agent, the 'underdetermined' agent and the weak-willed agent. But there are differences. We can understand these better if we see that the ordinary rational agent really is the 'underdetermined' agent. It just does not seem correct to say that an all-things-considered reason is directly motivational in the sense that the PPR view describe.

What we have found to be more plausible is the idea that there is more to an agent's intention to act than her judgement about how she should act (that is, her all-things-considered or endorsed reason for action). We also found that it is quite possible for an agent to intentionally act in a way that she did not intend to act. Not only is this possible but there are situations when this is rational as well. What emerged out of our discussion of these sorts of examples is a view of practical reasoning as a process with different stages and with the agent proceeding to make the transition from one stage to another before she finally acts.

There is, first of all, the evaluation stage during which the agent considers her reasons for action and comes to a conclusion about how to act. Before she acts however she must commit herself to her chosen course of action. To this end we discussed the transition from evaluative judgement to the intention to act in terms of the

agent becoming *committed* to her chosen action. This notion of commitment we elucidated in terms of the agent's favourable attitude towards her conclusion about how to act. I argued (and here I am influenced by Michael Bratman's view of intention) that in order to act on her evaluative judgement an agent must be disinclined to revise her deliberations. In other words if she intends to act then she treats her evaluative judgement as something not up for revision (unless there are new considerations of sufficient magnitude). We discussed this in terms of having a stable favourable attitude towards her evaluative judgement. This stable favourable attitude is her commitment to act, and this is what it means to intend to act. But the stability and the satisfaction that prevents the agent from revising too readily are not included in the evaluative judgement itself. We can say this because having an evaluative judgement does not prevent an agent from being weak-willed and it does not always determine the eventual action that will be performed.

So is it wrong to say that weakness of will is action knowingly performed against one's all-things-considered reason for action? I do not think it wrong to continue to describe weakness of will in this way. What we have learned however is that agents do not act on their reasons (that is, their evaluative judgements). Rather, they act on their intentions and intentions are only partly constituted by evaluative judgements. However it does make most sense for an agent to intend to act in accordance with her evaluative judgement. Any agent who failed to do this would be a sub-optimal practical reasoner. But in what sense can we say that this is *rationally* binding upon agents? I have argued that we should understand this process of moving from evaluative judgement to the intention or commitment to act not as guaranteed by the sheer logic of practical reason but rather as a regulative, norm-governed transition. Thus a weak-willed agent does perform an action that is against her all-things-considered reasons. There are two ways in which her weakness of will might manifest

itself. She might make a commitment that is in line with her evaluative judgement, but it might not be a stable commitment so she is weak-willed because she will not hold that commitment long enough for it to be motivationally efficacious. Or, alternatively she might form a stable commitment where that commitment is not based on the content of her evaluative judgement. This agent is weak-willed since she will go on to act on a commitment that she knows on reflection that she should not have formed.

What is interesting is that analysing this problem enables us to see that an agent's attitudes regarding her reasons for action play a crucial role in bringing about that action. I have argued that it is by examining cases of weakness of will that we see that reaching an evaluative judgement through practical reasoning is an incomplete account of how agents come to act. However this is not motivational scepticism – we can see that there are problems with the internalist assumption that endorsing a reason for action just is reaching an all-things-considered evaluative judgement. If we accept that there is a further transition between holding an evaluative judgement and endorsing the contents of the judgement we can make consistent the two intuitions we began with within a theory of practical reason. I have argued that an attitudinal account of commitment best explains the transition between evaluation and endorsement or intending, and the problem of weakness of will brings this to the fore.

But there is one conclusion that seems unavoidable, and that is, that an agent is not rationally bound to commit to acting on her evaluative judgement. Such an agent may be a critisably poor practical reasoner or we may say that she lacks *phronesis*, but she is not violating a law of reason. Thus an attitudinal account of practical reason does not fall prey to the charges of descriptivism and failure to account for the *authority* of all-things-considered reasons for action as levied by the PPR view. Weak-willed agents can be criticised for being sub-optimal practical reasoners on our alternative Reflective Dispositionalism account of practical reason. Satisfactorily

resolving the problem of weakness of will requires that we reject the claim central to the PPR view that having a reason for action (all-things-considered) is necessarily motivating.

## Bibliography


Allison, Henry (1990), *Kant's Theory of Freedom* (Cambridge: Cambridge University Press).

Anscombe, G.E.M. (1959), *Intention* (Oxford: Blackwell).
__ (1965), 'Thought and Action in Aristotle: What is 'Practical Truth'?' in *New Essays on Plato and Aristotle*, Renford Bambrough (ed.) (New York: Humanities Press).

Aristotle (1985), *Nichomachean Ethics,* ed. Terence Irwin (Indianapolis: Hackett). See also trans. by W.D. Ross (Oxford: Oxford University Press, 1980).

Audi, Robert (1990), 'Weakness of Will and Rational Action', *Australasian Journal of Philosophy* 68.
__ (1997), *Moral Knowledge and Ethical Character* (Oxford: Oxford University Press).

Austin, JL (1961), 'A Plea for Excuses', in *Philosophical Papers* (Oxford: Clarendon Press).

Baier, Annette (1991), *A Progress of Sentiments* (Boston: Harvard University Press).
__ (1996) *Moral Prejudices* (Cambridge, Mass.: Harvard University Press).

Baron, Marcia (1995), *Kantian Ethics Almost Without Apology* (New York: Cornell University Press).

Blackburn, Simon (1981), 'Reply to McDowell: Rule-Following and Moral Realism' in Holtzman and Leich (eds.) *Wittgenstein: To Follow a Rule* (London:Routledge)
__ (1984), *Spreading the Word* (Oxford: Oxford University Press).
__ (1992), 'Gibbard on Normative Logic', *Philosophy and Phenomenological Research* Vol.52.
__ (1993), *Essays in Quasi Realism* (Oxford: Oxford University Press).
__ (2000), *Ruling Passions* (Oxford: Oxford University Press).

Bratman, Michael (1979), 'Practical Reasoning and Weakness of the Will' *Noûs*.
__ (1987), *Intention, Plans and Practical Reason* (Cambridge Mass. :Harvard University Press).
__ (1999), 'Davidson's Theory of Intention', *Two Faces of Intention* (Cambridge: Cambridge University Press).
__ (1999), *Two Faces of Intention* (Cambridge: Cambridge University Press).
__ (2000), 'Reflection, Planning, Agency', *The Philosophical Review* Vol. 109, No.1.
__ (2003), 'A Desire of One's Own', *The Journal of Philosophy*, Vol.C, No.5.

Brink, David O (1989), 'Externalist Moral Realism', *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press).

Butler, Joseph (1953), *Fifteen Sermons Preached at the Rolls Chapel* ed. Revd. D. Matthews (London: Bell and Sons).

Charles, David (1984) *Aristotle's Philosophy of Action* (London: Duckworth).

Chisholm, R. (1964), 'The Ethics of Requirement', *American Philosophical Quarterly*.

Copp, David (1995), 'Moral Obligation and Moral Motivation', in Jocelyne Couture and Kai Nielsen eds., *On the Relevance of Metaethics: Canadian Journal of Philosophy* Supplementary Volume 21 (1995).

Cullity, Garret and Gaut, Berys (1997), *Ethics and Practical Reason* (Oxford: Oxford Clarendon).

Dancy, Jonathan (1993), *Moral Reasons* (Oxford: Blackwell).
__ (2000) *Practical Reality* (Oxford: Oxford University Press).

Darwall, Stephen (1983), *Impartial Reason* (London: Cornell University Press).
__ (1995), *The British Moralists and the Internal Ought 1640-1740* (New York: Cambridge University Press).
__ (ed.), (2003), *Virtue Ethics* (Oxford: Blackwell).

Davidson, Donald (1963), 'Actions, Reasons and Causes', *Journal of Philosophy* 60.
__ (1970), 'How is Weakness of the Will Possible?', in Joel Feinberg (ed.), *Moral Concepts* (New York: Oxford UP). Both repeated in Davidson *Essays on Actions and Events* (Oxford, 1980).
__ 'Paradoxes of Irrationality' (1982) in Wolheim and James Hopkins (eds.) *Philosophical Essays on Freud* (Cambridge: Cambridge University Press).
__ (1986), 'Deception and Division' in Jon Elster (ed.) *The Multiple Self* (London: Cambridge University Press)

Dunn, Robert (1987), *The Possibility of Weakness of Will* (Indianapolis: Hackett).

Elster, John (1986), 'Introduction' in *Rational Choice*, J. Elster (ed) (Oxford: Blackwell).

Falk, W.D. (1986), *Ought, Reasons and Morality* (Ithaca: Cornell University Press).

Fate-Norton, David (ed.) (1993), *A Cambridge Companion to Hume* (Cambridge: Cambridge University Press).

Frankena, William (1976), 'Obligation and Motivation in Recent Moral Philosophy' in K.E. Goodpaster ed., *Perspectives on Morality: Essays by William K. Frankena* (London: University of Notre Dame Press).

Frankfurt, Harry (1971), 'Freedom of the Will and the Concept of a Person', *The Journal of Philosophy*, Vol.LXVIII.
__ (1988), *The Importance of What We Care About* (New York: Cambridge).
__ (1999), *Necessity Volition and Love* (New York: Cambridge).

Gibbard, Allan (1985), 'Moral Judgement and the Acceptance of Norms', *Ethics* Vol.96.
__ (1990), *Wise Choices, Apt Feelings* (Oxford: Clarendon Press).
__ (1999), 'Morality as Consistency in Living', *Ethics* Vol.110.

Gilead, Amihud (1999), 'How is Akrasia Possible After All?' *Ratio*, Vol. XII.

Ginet, C. (1990), *On Action* (Cambridge: Cambridge University Press).

Goldman, Alvin (1970), *A Theory of Human Action* (New Jersey: Prentice-Hall).

Goodpaster, K.E. ed., (1976), *Perspectives on Morality: Essays by William K. Frankena* (London: University of Notre Dame Press).

Gosling, Justin (1990), *Weakness of the Will* (London: Routledge).

Guyer, Paul (ed.) (1998), *Kant's Groundwork of the Metaphysics of Morals* (Oxford: Rowman and Littlefield).

__ (2000), *Kant on Freedom, Law, and Happiness* (Cambridge: Cambridge University Press).

Hampton, Jean E. (1998), *The Authority of Reason* (Cambridge: Cambridge University Press).

Hare, R.M. (1963), 'Backsliding', *Freedom and Reason* (London: Oxford University Press).

Hill jr, Thomas E. (1998), 'Kant's Arguments for the Rationality of Moral Conduct' in Paul Guyer (ed.) *Kant's Groundwork of the Metaphysics of Morals* (Oxford: Rowman and Littlefield).

Holton, Richard (1999), 'Intention and Weakness of Will', *The Journal of Philosophy,* vol. XCVI.

Hume, David (2002), *An Enquiry Concerning the Principles of Morals,* (ed.) Tom Beauchamp (London: Oxford).

__ (2002) *A Treatise of Human Nature* (eds.) David Fate Norton and Mary J. Norton (London: Oxford).

Jackson, Frank (1984), 'Weakness of Will', *Mind.*

Johnson, Robert (1999), 'Internal Reasons and the Conditional Fallacy', *The Philosophical Quarterly,* 49.

Johnston, Mark (1989), 'Dispositional Theories of Value', *PAS* Supplementary Vol. 139-74.

Kant, Immanuel (1956) *Groundwork of the Metaphysics of Morals* (trans.) H. J. Paton (New York: Harper and Row).

__ (1992), *Critique of Pure Reason* (trans.) Norman Kemp Smith (London: MacMillan).

Kenny, Anthony (1966), 'The Practical Syllogism and Incontinence', *Phronesis* 11, 163-84.

Korsgaard, Christine (1986), 'Skepticism about Practical Reason', *The Journal of Philosophy,* , 24-6.
__ (1996) *The Sources of Normativity* (Cambridge: Cambridge University Press).
__ (1997) 'The Normativity of Instrumental Reason' in Cullity and Gaut (eds.) (1997).

Kydd, Rachel (1946), *Reason and Conduct in Hume's Treatise* (Bristol: Thoemmes Press).

Kubara, Michael (1975), 'Acrasia, Human Agency and Normative Psychology' *Canadian Journal of Philosophy* 5.

Melden, A.I. (1961), *Free Action* (London: Routledge and Keegan Paul).

Mele, Alfred (1986), 'Is Akratic Action Unfree?' *Philosophical and Phenomenological Research* 46.
__ (1987), *Irrationality* (London: Oxford University Press).
__ (1992), *Springs of Action* (Oxford: Oxford University Press).
__ (1995), *Autonomous Agents* (Oxford: Oxford University Press).
__ (2003), *Motivation and Agency* (Oxford: Oxford University Press).

McCann, Hugh (1972), 'Is Raising One's Arm a Basic Action?' *Journal of Philosophy* 4 89: 235-249

McDowell, John (1978), 'Are Moral Requirements Hypothetical Imperatives?', *Proceedings of the Aristotelian Society* Supplementary Vol. pp.13-29.
__ (1979), 'Virtue and Reason', *The Monist*, pp.311-350.
__ (1985), 'Values and Secondary Qualities', in *Morality and Objectivity* T. Honderich (ed.) (Boston: Routledge and Keegan Paul).
__ (1981), 'Non-Cognitivism and Rule-Following' in Holtzman and Leich (eds.) *Wittgenstein: To Follow a Rule* (London: Routledge)

Nagel, Thomas (1970), *The Possibility of Altruism,* (Oxford: Clarendon Press).
__ (1986), *The View from Nowhere* (New York: Oxford University Press).
__ (1997), *The Last Word* (London: Oxford).

Nozick, R. (1993), *The Nature of Rationality* (Princeton: Princeton University Press).

Nussbaum, Martha (1978) 'Practical Syllogisms and Practical Science' in her *Aristotle's "De Motu Animalium"* (Princeton: Princeton University Press).

Pears, David (1982), 'How Easy is Akrasia?' *Philosophia* 11.
__ (1984), *Motivated Irrationality* (Oxford: Clarendon).

Pettit, Philip and Smith, Michael (1993), 'Practical Unreason', *Mind*, 53-79.

Platts, Mark (1979), *Ways of Meaning* (Routledge and Kegan Paul).
__ (1981), 'Moral Reality and the End of Desire' in Mark Platts ed., *Reference, Truth and Reality* (Routledge and Kegan Paul, 1981).

Plato (1956), *Protagoras* and *Meno* (trans.)W. K. C. Guthrie (London: Penguin).
__ (1974), *The Republic* (trans.) Desmond Lee (London: Penguin).

Pugmire, David (1982), 'Motivated Irrationality Part I', *Proceedings of the Aristotelian Society,* 56.

Rapaport, Elizabeth (1973), 'Explaining Moral Weakness' *Philosophical Studies* 24.

Rawls, John (2000), *Lectures on the History of Moral Philosophy,* Barbara Herman (ed.) (Cambridge, Mass.: Harvard University Press).

Richman, Robert J. (1980), 'Acrasia and Practical Reasoning' *Pacific Philosophical Quarterly* 61.

Ridge, Michael (1998), 'Humean Intentions' *American Philosophical Quarterly* 35.

Rorty, A.O. (1980), 'Where does the Akratic Break take Place?' *Australasian Journal of Philosophy* 58.
__ (1980) *Essays on Aristotle's Ethics* (London: University of California Press).

Ryle, Gilbert (1949), *The Concept of Mind* (London: Hutchinson).

Santas, Gerasimos (1969), 'Aristotle on Practical Inference, the Explanation of Action, and Akrasia', *Phronesis* 14, 162-89.

Sayre-McCord, Geoffrey (1988), *Essays on Moral Realism* (New York: Cornell University Press).

Schneewind, J.B. (1998), *The Invention of Autonomy* (Cambridge: Cambridge University Press).

Sellars, Wilfrid (1970), 'On Knowing the Better and Doing the Worse', *International Philosophical Quarterly*, 10.
Sharpe, R. A. (1990) *Making the Human Mind* (London: Routledge).

Smith, Michael and Petit, Philip (1993), 'Practical Unreason', *Mind*, 53-79.
Smith, Michael (1994), *The Moral Problem* (London: Blackwell).

Stocker, Michael (1979), 'Desiring the Bad: An Essay in Moral Psychology' *Journal of Philosophy.*
__ (1990), 'Akrasia: The Unity of the Good, Commensurability, and Comparability' from *Plural and Conflicting Values* (Oxford: Clarendon).

Taylor, CCW (1980), 'Plato, Hare and Davidson on *Akrasia'* Mind 89.

Thalberg, I. (1977), *Perception, Emotion and Action* (Oxford: Basil Blackwell).

Thomas, Geoffrey (1993), *An Introduction to Ethics: Five Central Problems of Moral Judgement* (Cambridge: Hackett).

Thomson, J. (1977), *Acts and Other Events* (Ithaca, NY: Cornell University Press).

Tiberius, Valerie (2000), 'Humean Heroism' *Pacific Philosophical Quarterly* 81.

Watson, Gary (1977), 'Skepticism about Weakness of Will' *Philosophical Review* 86.
__ (ed.) (2003), *Free Will* (Oxford: Oxford University Press).

Walker, Arthur F. (1989), 'The Problem of Weakness of Will', *Nous*, 23.

Wallace, R. Jay (1990), 'How to Argue about Practical reason' *Mind*, Vol.99.

Wiggins, David (1980), 'Deliberation and Practical Reason' in A.O. Rorty (ed.) (1980).
__ (1980), 'Weakness of Will, Commensurability, and the Objects of Deliberation and Desire' in A.O. Rorty (ed.) (1980).

Williams, Bernard (1981) *Moral Luck* (Cambridge: Cambridge University Press).