Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXXII No. 1, January 2011

Resisting 'Weakness of the Will'

NEIL LEVY
Florey Neuroscience Institutes and Oxford Centre for Neuroethics

Weakness of the will is, apparently, an almost ubiquitous feature of everyday life. Yet it is also a deeply puzzling phenomenon. If an agent sincerely judges that it would be best to perform a particular action, and believes that they are able to perform that action, it is mysterious why – and how – they might intentionally perform an alternative action they take to be less preferable. In the face of this puzzle, it is tempting to be skeptical, and conclude that agents' avowals that some of their intentional actions conflict with their judgments as to how they (all things considered) ought to act are insincere. This paper defends a different kind of skepticism about weakness of the will. I deny neither that there is a phenomenon to be explained, nor that the phenomenon centrally involves the loss of self-control. The behavior we describe as weak-willed is real enough. But, I will argue, weakness of the will is not a *psychological kind*. That is, the folk psychological notion of weakness of the will, the notion imported largely uncritically into philosophy, does not correspond to a useful explanatory category for psychology. Instead, it is an instance of a broader phenomenon: agents switching from an effortful and basically rational mode of information processing to a more intuitive mode. The *concept* of weakness of the will, I will suggest, is useful neither for the explanatory purposes of psychology nor for the practical purposes of enhancing our ability to pursue the goals we value. But the broader phenomenon of which it is an instance is useful for both purposes. Hence we ought to abandon the narrower concept in favor of the broader.

I shall approach the broader phenomenon through the narrower. The account of the phenomenon I will offer will be empirically driven. I will argue that the empirical data give us powerful arguments in favor of certain views in the existing weakness of the will debate and against others. But I will go on to suggest that explaining the entire range of

data, experimental and observational, requires us to abandon the belief/desire framework within which existing accounts are elaborated, in favour of an account that adopts the vocabulary of cognitive psychology.

Weakness of Will: Theoretical Options

Weakness of the will occurs when an agent performs an intentional action despite believing that an incompatible action is both open to them and all things considered preferable. This working definition of weakness of the will leaves many questions open; for instance, whether the judgment that the unperformed action is preferable must be occurrent for the action to count as weak-willed. These are matters that a satisfactory account of weakness of the will should settle; it would beg substantive questions to build a resolution to them into the definition.

Accounts of weakness of the will must answer two major questions, which we may call the 'what' question and the 'how' question. The 'what' question concerns what psychological or mental states or entities must be postulated in order to explain weakness of the will. There are three main options available, each of which augments the next. Adopting Richard Holton's (1999) terminology, we may call them the basic Humean account, the augmented Humean account, and the will-power account. The Humean account explains weakness of the will in terms of the interaction of beliefs and desires. The augmented Humean account explains weakness of the will in terms of the interaction of beliefs, desires, and *intentions*, where intentions are irreducible to beliefs and desires. Finally, the will-power account – Holton's own view – postulates the existence of a separate faculty of the will, whose job it is to maintain our resolutions by keeping us from reconsidering them.

The 'how' question asks how the mental states or entities required to explain weakness of the will actually cause weak-willed behavior. There are two basic options here: judgment-based and desire-based accounts. Traditional judgment-based accounts typically explain weakness of the will by distinguishing between the (implicit) judgment which causes the actual action and the (explicit) judgment which the agent expresses. Accounts of this sort date back to Aristotle. More recent judgment-based accounts have been proposed by Davidson (1970) and Tenenbaum (2007). The distinction between two kinds of judgment, the agent's implicit and her explicit judgment (in Davidson's terms, her *all things considered* judgment and her *unconditional* judgment; on Tenenbaum's account, her *direct* and *oblique* cognitions)

is invoked to explain the discrepancy between what the agents *says* and what she *does*.

Desire-based accounts assimilate weakness of the will to something akin to compulsion. On these accounts, agents act against their resolutions because their desires impel them to. Such accounts face an obvious objection: if weak agents are caused to act as they do by their desires, what grounds do we have for distinguishing weakness of the will from compulsion? There are several responses to this objection in the recent literature. Watson (1977) argues that the distinction between compulsive actions and weak actions is normative: we describe agents as weak when they are overcome by desires they *ought* to have been able to resist. Smith (2003) argues that there is a sense in which weak agents *could have* resisted their recalcitrant desires: weak agents, unlike compulsive agents, possess the rational capacities to act as they ought. Finally Mele (1995) argues that weak agents differ from compulsive agents, *inter alia*, in that there are typically things they can do, at the very moment of action, to reduce the causal force of their recalcitrant desires.

This brief sketch of the some of the landscape of the weakness of the will debate serves as the background against which we must understand the significance of the empirical evidence on self-control failures. In some ways, this evidence speaks directly to the debate as it has been framed in the recent philosophical literature, supporting a judgment-based account of weakness of the will over a desire-based account. In other ways, however, it provides us with reasons to abandon much of the traditional framework, in favour of the concepts and vocabulary of cognitive psychology.

The Empirical Evidence

I shall briefly sketch the main thrust of the empirical evidence; further details will be added as they become relevant. There is now a substantial body of work on what has become known (somewhat unfortunately, I think) as *ego depletion* (Baumeister et al. 1998; Baumeister 2002). Ego depletion refers to the depletion of an energy source preferentially drawn on by self-control mechanisms. The classic research on ego depletion proceeds as follows: subjects are divided into two groups: an ego depletion group and a control group. The depletion group is given a task that draws upon their self-control reserves – say, watching a funny movie without smiling – while the control group is given a task matched for effortfulness but which does not require much self-control – say, rating various options in terms of desirability. Then both groups are given a common task that requires self-control: for example holding one's hand in icy water (the 'cold pressor task') or attempting to solve

an anagram puzzle that is in fact insoluble. The finding is that subjects in the ego depletion group persist a significantly shorter time at the self-control task then subjects in the control group. These results seem to indicate that self-control resources are temporarily depleted when they are drawn upon, and that when self-control reserves are low, engaging in tasks that require self-control becomes much more difficult.

Clearly, ego depletion is of great philosophical interest. But is it a good model for weakness of the will? Obviously, subjects in these experiments do not have a prior commitment to persisting at the cold pressor task or the other tests used to measure ego depletion. It may therefore be doubted whether they exhibit any weakness of the will at all. This doubt can be allayed: several ego depletion experiments have examined the behaviour of subjects with pre-existing commitments to self-control in a particular domain. Vohs & Heatherton (2000) measured the consumption of ice cream in subjects ostensibly engaged in a flavour rating exercise. Ego depleted chronic dieters ate significantly more than ego depleted non-dieters as well as non-depleted dieters. Kahan et al. 2003 measured the consumption of cookies of depleted dieters in what was ostensibly a taste perception test; once again, 'restrained' eaters consumed significantly more. Restrained eaters come to the experimental situation with a pre-existing resolution; the ego depletion paradigm led them to act contrary to it. Thus, ego depletion appears to generate weakness of the will.

Why conclude, from the fact that restrained eaters subject to ego-depletion eat more than either ego depleted non-dieters or non-depleted dieters, that ego depletion generates weakness of the will? Restrained eaters are subjects with a prior commitment to limiting their intake of food, and especially of high-calorie foods such as cookies and ice cream. Hence their consumption of the tempting foods conflicts with their prior resolutions. On some accounts, this behaviour just *is* weakness of the will (Holton 1999; McIntyre 2006; Dodd forthcoming). Philosophers who take this line distinguish between *akrasia*, understood as action that conflicts with the agent's concurrent judgment, and weakness of the will proper, understood as action that conflicts with a prior intention or resolution. Clearly there are two distinct phenomena here, as evidenced by the fact that they can dissociate: *akrasia* (so defined) is neither necessary nor sufficient for weakness of the will. It is not necessary, because agents can resolve to act in ways that conflict with their judgments and subsequently exhibit weakness of the will by failing to carry through their resolutions; it is not sufficient because agents can fail to act in accordance with their judgments while failing to act weakly (when, for instance, they are not at all motivated to act in accordance with a judgment). Since only actions that conflict with an

intention or resolution are properly described as weak, philosophers who have taken this line suggest, we should identify weakness of the will with these kinds of actions and not with *akrasia*. If we are convinced by these arguments, we ought to describe the behaviour of the subjects in the above experiments as weak-willed, and conclude that ego depletion generates weakness of the will.

It is open to us, of course, to reject the claim that weakness of the will ought to be identified with action that conflicts with agents' prior resolutions or intentions. We might insist that weakness of the will requires, or is also instantiated, when an agent acts in a way that conflicts with their current judgment, whether or not it also conflicts with a prior resolution. As we shall see, there are good reasons to think that the subjects in ego depletion experiments do not exhibit weakness of the will if it is best understood as action that conflicts with a simultaneous occurrent judgment. However, there are good grounds, philosophical and empirical, for doubting that weakness of the will requires a simultaneous occurrent contrary judgment.

Some philosophers appeal to introspection to establish that weakness of the will requires a contrary occurrent judgment. FitzPatrick (2008), for instance, points out that he often knows perfectly well that he ought not to be behaving as he is, *even as* he digs into a bowl of chocolate ice-cream. But this claim, which seems plausible, does not establish the *occurrent* judgment model of weakness of the will. I may know things dispositionally without forming the correlative occurrent judgment. Moreover, if I resolve at t not to ϕ, and I remain committed to the resolution at $t1$, I may rightly attribute to myself the dispositional belief that I ought not to ϕ for the interval between t and $t1$, even if I acted in a way that conflicted with my resolution during that interval. Looking back on our behaviour, we may therefore follow FitzPatrick and truthfully say that even as I acted I *knew* I shouldn't. So long as the knowledge is dispositional, such a claim is entirely compatible with the resolution model of weakness of the will.

Perhaps philosophers like FitzPatrick think that introspection shows that we sometimes occurrently believe that we ought not to act as we do when we exhibit weakness of will. Certainly agents often are (occurrently) aware of their resolutions when they act weakly, but that fact is not sufficient to attribute to them the occurrent wholehearted judgment that they ought not to perform the act they perform. They may simply make an exception of the current occasion. The claim that agents occurrently and wholeheartedly believe that they ought, on this very occasion, not to act as they do is an empirical claim. So far as I know, there is no decisive evidence against it, but there is strong evidence for the unreliability of introspective claims like it. In general, social and

cognitive psychology have shown that we have less secure access to our mental states and motivations when we act than we commonly think (Schwitzgebel 2008). Given that this is the case, we ought to place little weight on the introspective claim. Alternatively, perhaps proponents of desire-based accounts are advancing a conceptual claim: perhaps they think that our concept of weakness of will *requires* that the action conflicts with an agent's occurrent judgment. If an agent acts in a way that conflicts with a resolution, they might maintain, they act weakly *only if* they also occurrently judge that they ought not so to act; otherwise, they have simply changed their mind. Since changing one's mind does not necessarily involve any failure of practical rationality, it cannot serve as a criterion for weakness of the will.

This is a challenge we ought to take seriously. At very least, proponents of judgment-based models owe us an explanation of how to distinguish weakness from *mere* changes in mind. But I do not believe that meeting this obligation is very difficult. Genuinely weak judgment-shifts are relatively brief and transitory; once the depleted resources are restored, the agent typically regrets the action. Changes of mind are long-lasting and induce stable states in agents; typically (though not necessarily) they have as their cause deliberation by the agent and not ego depletion. It is precisely because changes of mind are produced through deliberation and are not regretted by agents that we – rightly – do not regard them as involving failures of practical rationality. Ego depletion induced weakness is very different. It reflects a failure by the agent to control their mental life, and is therefore appropriately regarded as a failure of practical rationality.

Our psychological concepts ought to be sensitive to the actual empirical data. Since the data demonstrate that behaviour which is very like weakness of the will involves actions that conflict with resolutions, we should identify weakness of the will with such behaviour, until there are good grounds for revising our view. The fact that there are independent philosophical arguments, developed by Holton (1999) and McIntyre (2008), for the same conclusion only serves to buttress the case. Since ego depletion does therefore appear to generate weakness of the will, I will use the phenomenon as a model for weakness of the will, attempting to illuminate the latter in light of evidence drawn from studies of the former.[1]

[1] One caveat is in order: it does not follow from the fact that weakness of the will is often or typically caused by ego depletion that all instances of weakness of the will have the features I will outline. However, until there is convincing evidence that there are cases of weakness of the will that are not caused by ego depletion, we are justified in treating ego depletion as a model for *the* phenomenon.

We can learn a great deal about weakness of the will from the study of ego depletion. The first thing we can learn is that a judgment-based model of weakness of the will seems to be correct (though a judgment-based model that differs significantly from the traditional one). There are two sets of evidence supporting this conclusion. The first set comes from experiments which ask ego depleted subjects to choose options to be delivered at some specific time in the future – say, asking them on a Wednesday to select a film to be watched on the weekend. Depleted subjects select fewer highbrow films than non-depleted (Wang et al. under submission). This behaviour supports the claim that ego depletion causes weakness of the will by altering agents' judgments about how they ought to behave. Obviously, this claim is true only if the subjects in this experiment actually exhibited weakness of the will, and if the best explanation for their weakness of the will involves a shift in their judgments. I take these questions one at a time.

Why think that the subjects in these experiments exhibited weakness of the will? Recall, first, that we have already established that ego depletion is a good model for weakness of the will. If we have good reason to attribute to subjects in the experiment appropriate prior mental states, then we can conclude that the effect of the experimental manipulation was to induce weakness of the will. There are two mental states that must be appropriately attributed to the agents. First, it must be the case that they have a prior intention, resolution, or related attitude to the effect that they ought to choose highbrow films rather than trashier films; second, they must find the trashier films tempting (recall our evidence from the diet paradigm: only subjects who find food tempting consume more after depletion). There is indirect but persuasive evidence for a prior standing belief that highbrow films were seen as more choiceworthy than trashy films. It consists in the fact that the control group did choose significantly more highbrow films than the ego depletion group. Given that the subjects came from similar demographic backgrounds, and were randomly assigned to one group or the other, the best explanation of the difference in behaviour is that the ego depletion led subjects to exhibit weakness of the will. It is this very fact – the fact that we have good reason to attribute to the subjects a (normally distributed) belief that highbrow films were more choiceworthy, but the experimental manipulation induced a difference in behaviour – that provides us with the evidence for the second claim; that subjects found trashier films tempting. In isolation, the experiment might be susceptible to a variety of alternative explanations, but in the context of the body of

evidence from many other closely related experiments, we ought to conclude that the subjects exhibited weakness of the will.

Turn, then, to the second claim: that the judgment-based model best explains the mechanisms at work in the weakness exhibited here. The evidence consists, primarily, in the fact that the desire-based and judgment-based explanations seem to yield contrasting predictions regarding how subjects will make future oriented choices. Desire-based models of weakness of the will predict that agents will be overcome by the immediate attractiveness of available goods, but, given that the agent continues to occurrently believe that other goods are preferable all things considered, they seem to predict that the subjects will make future-oriented choices in line with their judgments. Since this prediction was falsified by the behaviour of subjects, we ought to reject desire-based models in favour of judgment-based competitors.

Let me expand on this claim. All the available evidence, from animal studies and human studies alike, suggests that immediately available rewards work far more powerfully on our desires than do distant rewards, and that the choice of distant rewards which conflict with one's resolutions reflects a change in judgment: when subjects sincerely judge, at *t*, that they ought to perform some act at a future time *t1*, they are (in the absence of some incapacitating condition) capable of putting that judgment into effect at *t,* minimally by expressing it. Consider the evidence from hyperbolic discounting (Ainslie 2001). Hyperbolic discounters, both human and animal, judge at *t* that they ought to act in one way, yet at *t1* they find themselves judging in a way that conflicts with that judgment. For instance, they might judge that they ought to refrain from consuming some desirable good, at *t*, but when the opportunity to consume the good is available, they act on it. This suggests that temporal distance significantly reduces the effects of the desirability of goods on our behaviour: Even pigeons who learn that they are subject to hyperbolic discounting, such that they are unable to resist consuming a small reward when it is available despite the fact that could they wait they would receive a larger later reward, are able to learn to use commitment devices to maximize their rewards. By pecking a button before the smaller reward becomes available, they prevent themselves consuming the small reward, thus forcing themselves to wait for the larger later reward (Rachlin 2000). The immediate availability of a reward acts powerfully on agents (whether by altering their preferences or compelling them to act against their preferences); whereas temporal distance to reward reduces its power greatly. Hence, when agents express a preference for a temporally distant good, we

have good evidence that this expressed preference really reflects their judgment at the time they make it.[2]

On a desire-based account, the desires should therefore work far more powerfully on immediate rewards than on distant. It is therefore implausible to think that when agents choose a film for a relatively distant future occasion, they are overcome by a desire to choose against their own better judgment. Given that the reward is deferred, it does not seem to have the causal power that comes from its promise of immediate sensuous gratification. Hence the pattern of future choices exhibited by ego depleted subjects seem to reflect what Holton (1999) calls judgment-shift; a temporary reversal of the agent's preferences. That is not to say, of course, that the rewarding characteristics of tempting goods do not play a causal role in explaining agents' choices; it is only to suggest that the causal route from these characteristics runs through the agents' judgments as to what it would be best to do.

The second line of evidence for the claim that ego depletion causes weakness of the will via a change in judgment, not via the influence of a desire over an agent who continues to make judgments in line with their resolutions, comes from studies focusing directly on the attitudes of ego depleted subjects. Wheeler et al. (2007) gave subjects counterattitudinal arguments. Some of these arguments were designed to be strong, some weak. Depleted and non-depleted subjects were equally convinced by strong arguments, but depleted subjects were significantly more convinced by weak arguments. What explains this effect? The experimenters asked subjects to estimate the degree of their attentiveness to the message and of the effort exerted in assessing it, and found no significant differences between depleted and non-depleted subjects. The increase in persuasiveness for depleted subjects of weak arguments does not seem to be due to these factors. Instead, ego depletion seems to lead to lower quality processing of the message content. There is strong evidence that accepting truth claims is the cognitive default

[2] It should be noted that Ainslie does not interpret the scope of the judgment-shift seen in hyperbolic discounting in the way I have suggested. He claims, rather, that hyperbolic discounters exhibit an indexical preference shift: when the time for consumption is imminent, they shift from judging they ought to refrain to judging that they ought to refrain *on every occasion except the present*. The evidence from Wang et al. seems to conflict with this claim: under the influence of temptation, the subjects shifted from preferring that they watch worthy movies now and in the future to preferring to watch trashy movies now and in the future. This issue requires further exploration, given that the kind of reasoning that Ainslie describes, in which a subject does not alter their view about how they ought to act in general but decides to make an exception of the current occasion, *seems* common. If Ainslie were to prove right, and the scope of judgment-shift was (typically) restricted to the current occasion, the behaviorial evidence would favor neither the desire-based nor the judgment-based account; it is the broad scope of the expressed preference that constitutes evidence in favor of the latter.

position. This is the result of the apparent fact that to understand a proposition is to take it to be true, if only momentarily (Gilbert 1991). Thus, we comprehend a negative claim by imagining the conditions that must obtain for the claim to be true, and only subsequently inserting a mental negation sign in front of it (part of the evidence for this claim comes from studies of processing speeds; it takes longer to process a negative claim than a positive). In the absence of the processing resources needed to retrieve or generate contradictory information and apply it to the message content, the default tendency to acquiescence takes over and we accept the message as true. This hypothesis is supported by the fact that depleted subjects in Wheeler et al.'s study reported significantly more positive thoughts while assessing weak counterattitudinal arguments than did non-depleted subjects.

This suggests a – judgment-based – mechanism explaining the occurrence of weakness of the will. In response to temptation, subjects spontaneously generate or retrieve from memory arguments in favour of weak-willed action. Since they lack the cognitive resources to reject these arguments, they experience judgment-shift. They come to judge that the benefits of succumbing to temptation are higher, or the costs of giving in lower, or both, and act accordingly. The suggestion that this mechanism explains weakness of the will is plausible only if it is true that generating or retrieving arguments in favour of acquiescence is less effortful than assessing these arguments. Why should that be the case? Temptations, I suggest, automatically generate arguments in their favour. They might even be said to *constitute* arguments in their favour: for typical temptations, the major argument in their favour is that consumption of the tempting good is pleasurable. Ego depleted individuals exhibit a stronger preference for the affective properties of products than do non-depleted (Baumeister 2008), suggesting a greater susceptibility to such pleasures. But there is another route whereby arguments in favour of succumbing might effortlessly be generated: simply by retrieval from memory. Typically, we experience weakness of the will with regard to goods whose attractions we know all too well; often they are goods we have antecedently resolved to resist. We are therefore likely to have considered arguments in favour of giving in to temptation, and be able to retrieve them without effort. It is worth noting that rote memory is not affected by ego depletion (Schmeichel, Vohs, and Baumeister 2003).[3]

[3] Obviously, this suggestion is plausible only if we can explain an asymmetry in memory retrieval: why should the subject be able effortlessly to retrieve arguments in favor of consumption, but not arguments in favor of maintaining the resolution? Since (as we shall soon see), ego-depletion tends to switch the subject to system 1, he or she will engage in a biased memory search, of the kind we see exemplified in the confirmation bias (Nickerson 1998); a paradigmatically system 1 process.

Beyond Weakness of the Will

So far, we have suggested that the evidence from ego depletion gives us reason to favour one of the competing accounts of weakness of the will, a judgment-based account, over its desire-based rival. Does it provide us with grounds to settle the other debate, concerning which mental states or processes we need to postulate to explain the phenomenon? Holton (2003) has argued that it does: the data support the existence of a separate faculty of will-power, and therefore demonstrate the inadequacy of a Humean or an augmented Humean account. Holton suggests that the empirical evidence demonstrates that weakness of the will is a broader phenomenon than rival accounts can explain; hence it supports the existence of a faculty – willpower – that is as broad in its function as the domain of self-control. I will argue that Holton is right in thinking that the breadth of the phenomena is the key to their explanation, but that he underestimates this breadth. The causes and effects of ego depletion are broader than his account can explain; so broad that we are required to jettison the notion of weakness of the will as a psychological kind altogether.

On the Humean and augmented Humean accounts, weakness of the will (*all* intentional action) is explained in terms of the causal force of the interacting elements; on the Humean account these elements are beliefs and desires, whereas the augmented Humean account adds intentions to the mix. Holton argues that these accounts cannot explain the systematic nature of the loss of self-control. The factors which undermine agents' ability to maintain their resolutions affect self-control *globally*. It is difficult to explain this fact within the framework of a simple belief/desire, or even belief/desire/intention framework. Why should *all* the desires we intend to resist – and only these desires – strengthen when we are ego depleted? It might be suggested that the result is due to the strengthening of basic appetites, our appetites for food, sex and other sensual pleasures. But this explanation does not fit the facts. Dieters who are depleted eat more; non-dieters do not eat more (stressed dieters eat more; stressed non-dieters actually eat less). It seems that the self-control needed to maintain resolutions is preferentially weakened by depletion, regardless of the content of the resolutions. Neither Humean account seems able to account for this fact. Holton takes this to be powerful evidence in favor of the existence of a faculty of will-power.

Yet if Humean accounts, even augmented by irreducible intentions, fail to explain the phenomenon of ego depletion, Holton's will-power account is equally at a loss to explain all the data. Ego depletion has broad and systematic effects, as Holton stresses. But these effects are far broader than he appreciates. Ego depletion does not preferentially

affect our resolutions or even our self-control; it affects a range of cognitive functions. In order to understand which functions are affected and how, we need to abandon, not tinker with, the Humean framework. We need to adopt the concepts and vocabulary of cognitive psychology.

Cognitive psychology often divides cognitive processes into two basic systems: system 1 and system 2. System 1 is the evolutionarily ancient system we share with many other mammals. It consists of a set of mechanisms that respond automatically to stimuli, without the need for oversight from consciousness. System 1 processes are fast, ballistic and undemanding of cognitive resources; they will do their thing, given their proprietary input, regardless of the availability of other mechanisms. They cannot be stopped from functioning or from producing their output, except by preventing them receiving their inputs. They are often – perhaps always – modular (Stanovich 1999). System 1 processes operate in parallel. System 2 has the opposite profile: it consists of mechanisms that are slow, operate serially rather than in parallel, and are demanding of cognitive resources. System 2 processes are rule-governed and conscious. System 2 is distinctive of human beings (whether or not it is unique to human beings) and makes possible our most remarkable cognitive achievements. But the overwhelming majority of our actions are produced by system 1 which takes care of our basic survival needs and much more besides.

System 2 never operates on its own (unlike system 1). Attention is directed toward features of the environment by system 1, presumably when conditions are such that the resource-intensive system 2 will typically improve the quality of the resulting response. Moreover, system 1 biases the responses of system 2 in various ways, by altering the weight of influences on system 2 processes. System 1 processes always stand ready to take over from system 2 processes. Since the latter are demanding of cognitive resources, agents often cannot implement or continue to implement them. When the agent is stressed, tired or under cognitive load (multitasking, for instance), system 1 processes pick up the burden. The effect is measurable: agents act less flexibly, and in ways that are more stereotypical when their responses are generated by system 1.

All of this should immediately make us think of the ego depletion paradigm. Self-control, too, is slow, demanding and draining of cognitive resources. It is weakened or lost under conditions which look for all the world like the conditions which make agents switch from system 2 to system 1. I think that the resemblance between ego depletion and the effects of cognitive load is not a coincidence: self-control is a system 2 process, and its loss switches us to system 1 (Baumeister et al. 2008).

There is plentiful evidence for this claim. Let's begin with the effects of ego depletion on paradigm system 2 processes, such as logical thought. The same tasks that deplete self-control resources also lower performance on IQ tests while leaving system 1 processes like rote memory unaffected (Schmeichel, Vohs, and Baumeister 2003). Choice is, by itself, ego depleting. Baumeister and Vohs (2007) report a study in which participants chose items for a bridal registry. There were two conditions: a four minute task and a twelve minute task. Participants who enjoyed the task showed no depletion after the four minute task, whereas those who disliked it found the task depleting. The twelve minute task was depleting for both groups of participants. Deliberating without choosing was also somewhat depleting, though not as depleting as choosing. Choosing is apparently ego depleting; moreover, the degree of depletion is a function of the difficulty of the choice (Baumeister et al. 2008). Choosing and ranking items are both tasks for which we ordinarily employ system 2.

As we saw above, ego depletion also affects our ability to intelligently process arguments. Ego depleted individuals are more likely to accept counterattitudinal arguments, even when these arguments are weak. All this evidence suggests that ego depleted individuals rely more on system 1 than on system 2. Hamilton, Hong and Chernev (2007) set out to test this hypothesis directly. They primed subjects to rely on system 2 (by having them perform a few mental arithmetic problems – too few to deplete system 2) or system 1 (by showing them a visual figure with two interpretations, like the Necker cube, and asking them which they saw first). A third group also performed mental arithmetic problems, but many more of them, in order to induce ego depletion. The depletion group and the system 1 group showed indistinguishable patterns of response in a test of consumer preferences. Further evidence comes from a test of the effect of ego depletion on the tendency of consumers to ignore irrelevant alternatives in making choices. Ordinary agents are susceptible to the asymmetric dominance effect, in which the presence of an irrelevant alternative influences which option is chosen. This effect is believed to reflect the operation of system 1 processes. Masicampo and Baumeister (2008) found that ego depletion increased susceptibility to the effect.

Neuroanatomical evidence supports the claim that ego depletion draws upon system 2. Self-control involves regions in the prefrontal cortex (Banfield et al. 2005). The prefrontal cortex is also the site involved in effortful controlled – system 2 – processes, such as decision-making and logical reasoning. The most parsimonious explanation of all the data, then, is that ego depletion is in fact the result of drawing down of system 2 resources, pushing the agent into system 1. The

evidence suggests that modular systems prompt the change from system 2 to system 1, in order to preserve sufficient system 2 resources to deal with future contingencies. Hence ego depletion is not the product of the exhaustion of system 2 resources, but of mechanisms for their conservation. Accordingly, motivating the subject – for instance using cash incentives or by prompting them to think about their values – can significantly affect the degree of depletion exhibited. Depletion is task-relative. However, these short-term antidotes which prompt the agent to switch back to system 2 leave her even more depleted than before (Baumeister et al. 2008).

Holton utilizes the ego depletion literature in arguing for the existence of a dedicated faculty of will-power, which functions to block reconsideration of our resolutions. The evidence from ego depletion that I have just reviewed suggests that he is wrong. The broad and systematic effects of ego depletion are not preferentially exhibited in the domain of self-control or the maintenance of resolutions at all; nor are the effects produced by temptation alone. Instead, ego depletion is caused by engagement in any of the much broader class of system 2 processes, which involve effortful cognitive processing, in particular, but not only, the inhibition of prepotent responses. Ego depletion is produced by Stroop tasks, in which subjects have to name the colors rather than reading the (conflicting) words (Webb and Sheeran 2003); Stroop tasks involve the inhibition of responses, but not temptation to break a resolution. It is also produced by having to make choices, which does not involve the inhibition of a response at all. It is even produced by exaggeration of prepotent responses (Schmeichel et al. 2006).

Thus, the most plausible explanation of ego depletion sees it as involving mechanisms and situations much broader than those at work in weakness of the will. Weakness of the will is simply a special case of a broader phenomenon. I have suggested that this broader phenomenon is the depletion of system 2 resources, throwing the agent back onto the more plentiful, but predictably inflexible, system 1 processes.

Here is a brief sketch of how weakness of the will may occur. Exposure to a temptation tends automatically to generate an argument or quasi-argument in favor of consumption, perhaps as a response to the affordances for the subject of the temptation. It takes effort to resist this temptation, by generating contradictory information and applying it to the content of the argument. Sometimes agents are unable to exert the effort needed for maintaining their resolutions. Sometimes they lack the system 2 resources to generate the contradictory information or apply it to the argument, due to recent demands on system 2; sometimes they have sufficient resources to resist for a while, but as the temptation persists these resources are drained. At some point,

modular (system 1) processes which have the job of conserving system 2 resources for unexpected contingencies assess the resolution as insufficiently important to justify expending further resources on it, and switch off the tap. At this point agents switch from system 2 to system 1, which outputs the judgment that giving in has greater benefits than costs. The agent experiences judgment-shift and acts accordingly.

As mentioned above, this story depends for its plausibility on the association of the resolution with system 2 and the motivation in favor of succumbing with system 1; only on this condition is it plausible that maintaining resolutions is more effortful than succumbing and that depletion of system 2 leads to indulgence via judgment-shift. Many resolutions have the aim of restricting the satisfaction of basic appetites, for food and other intrinsically rewarding activities and goods; with regard to resolutions with these kinds of contents, the association between the resolution and system 2, on the one hand, and the temptation and system 1, on the other, is obvious. But ego depletion seems to weaken our resolve to maintain our resolutions, irrespective of their content. How do resolutions come to be associated with system 2, and temptations with system 1, regardless of content?

If temptations automatically and effortlessly generate arguments in favor of a course of action, then when we are ego depleted we will be biased toward accepting that argument. Recall the evidence that ego depleted individuals tend to accept weak arguments, even when they are counterattitudinal (Wheeler et al. 2007). System 1 might typically be biased toward propositions that are ego-syntonic or otherwise gratifying, but it appears that the bias toward acceptance is stronger, strong enough to trump the former bias.

Skepticism About Weakness of the Will

These considerations give us a powerful reason to doubt the existence of weakness of the will, understood as a discrete phenomenon. The skepticism about weakness of the will motivated by the evidence presented differs from more traditional skepticism inasmuch as it denies neither that agents sometimes act against their resolutions, nor that so acting involves a loss of self-control of some sort. Instead, the skepticism concerns the existence of weakness of the will as what we might call a *psychological kind*. The behavior we call weakness of the will exists, but the concept is useful neither for the explanatory purposes of psychology, nor for the practical purposes of increasing our ability to maintain self-control.

The psychological causes of weakness of the will are not restricted to the domain of resolutions and their maintenances, or temptations

more generally. Temptations do deplete our resources and make us vulnerable to subsequent succumbing. But a range of psychological and, for that matter, physiological phenomena are also depleting of the resources required for self-control. Ego depletion is caused by choices or by engaging in effortful thought, for example mathematical calculations. Ego depletion draws down the reserve of glucose dedicated to the energy-intensive prefrontal cortex; accordingly, any physiological process which decreases the availability of glucose affects our ability to maintain our resolutions: poor glucose tolerance, diabetes, hypoglycemia and poor diet all make us more vulnerable to temptation (Gailliot and Baumeister 2007). Correlatively, the effects of effortfully maintaining our resolutions are not restricted to the domain of resisting temptations; instead, they involve a more general tendency to switch us to system 1 processes. The effects are global, and involve a greater tendency to rely upon heuristics and other cognitive shortcuts, a decreased ability to reason logically, and even a decreased willingness to engage in helping behavior (Gailliot et al. 2007) and an increased propensity to cheat (Mead et al. 2009).

Since the causes and the effects of ego depletion cut across domains, rather than being limited to resistance of temptations, we should conclude that weakness of the will does not equate to a psychological kind, of the type that would figure in any science of the mind. Moreover, since we can best avoid losses of self-control by ensuring that we have sufficient resources to meet temptations when they arise, and so ensuring requires being attentive to the entire range of causes of ego depletion (as well as to the ways in which ego depletion can be compensated for; most simply, by ingesting glucose (Gailliot and Baumeister 2007)), the concept of weakness of the will has no useful role to play in guiding our everyday actions. But when a concept plays a useful role neither in science nor in guiding action, we have no reason to retain it. Instead, we should abandon it in favor of the broader phenomenon, the drawing down of system 2 resources.

Two Objections

Before concluding, let me briefly address two objections. The first is an empirically motivated objection to the main thesis, that weakness of the will is best understood as an instance of the more general phenomenon of agents switching from system 2 to system 1 processes; the second accepts the thesis but rejects the claim that this gives us a decisive reason to jettison the concept of weakness of the will.

The empirical objection is predicated on the observation that skill at self-control seems to dissociate from skill at system 2 processes

generally. If self-control just is the application of the domain-general system 2 to the problem of resolution maintenance, then we ought not to see any such dissociation (Richard Holton, personal communication). The dissociation may be cited as evidence in favor of Holton's view, according to which there is a discrete faculty of willpower, which has the role of preventing reconsideration of resolutions. Postulating the existence of such a faculty would help to explain why individuals differ from each other with regard to their self-control, independently of their differences in, say, logical reasoning.

The empirical evidence suggests that preventing ourselves from reconsidering our resolutions is indeed an effective way of avoiding succumbing to temptation, and no doubt we rely upon system 2 resources to prevent ourselves from reconsidering. But the most plausible explanation of how we avoid succumbing by way of avoiding reconsideration does not cite plentiful ego resources to explain how we prevent ourselves reconsidering; instead it explains our plentiful ego resources by citing our success in avoiding reconsideration. That is, successful delayers succeed not through strength of will, or large reserves of system 2 resources, but by avoiding taxing these resources. That is indeed how the subjects in Walter Mischel's studies, cited by Holton (2003) as evidence in favor of his view, succeeded in delaying gratification. When the children in these studies focused on the rewards for delay (a more desirable food object than the immediately available reward), they quickly succumbed. When they were given strategies to distract themselves (thinking fun thoughts or playing with a toy) they were able to delay very much longer than children given no such instruction. Because these children successfully distracted themselves, they did not need to tax their system 2 resources. People differ in their system 2 resources, but these differences do not explain their ability to delay gratification. Instead, it is their ability to deploy self-distraction techniques which explains self-control (Mischel notes that given the right instructions, 'virtually all subjects, even young children, could manage to delay for lengthy time periods' (1972: 217)).

If Mischel is indeed right, then the ability to delay gratification depends not on strength of will, or the state of system 2 resources, but on the ability to deploy a skill (Mischel and Mischel 1983). By avoiding reconsideration – by focusing on properties of a temptation that are not themselves rewarding, or thinking about something else altogether – those who exhibit strength of will are able to avoid taxing their system 2 resources. They utilize system 2 resources for this purpose, but directing one's thoughts away from a temptation is less demanding of these resources (when it is skillfully done) than is effortfully resisting the temptation. Note that this view is compatible with Mischel's

well-known claim that the ability to delay gratification is a stable trait that is a predictor of success in many arenas including academic success, social competence, attentiveness, concentration, the ability to form and execute plans and so on (Mischel, Shoda & Rodriguez 1989). Ability to delay gratification may be predictive either because children spontaneously learn to deploy the relevant skill, and learning to deploy the skill is itself predictive of success, or because measuring the state of system 2 resources is predictive of success.[4]

Even if the cogency of the argument in favor of understanding weakness of the will as an instance of the broader phenomenon of agents switching from system 2 to system 1 is accepted, the conclusion that we ought to jettison the concept of weakness of the will might be resisted. It is surely one thing to say that weakness of the will is an instance of a more general phenomenon, and quite another to say that we ought to abandon the concept altogether. Weakness of the will might not be the broadest explanatory concept available, but it might be useful to us, as agents, perhaps precisely because it is narrower. Weakness of will might relate to the broader phenomenon of agents switching from system 2 to system 1 in a manner analogous to the way in which Newtonian mechanics relates to general relativity. Newtonian mechanics is accurate enough to guide almost all of us under almost all conditions, and the costs in computational complexity that would be incurred were we to abandon it in favor of the more accurate theory would vastly outweigh the gains in accuracy.

But weakness of will does not relate to the system 1/system 2 framework in the way that Newtonian mechanics relates to relativity. The drawing down of system 2 resources by tasks that are not normally conceptualized as requiring willpower is ubiquitous in daily life. It is not just, and not even especially, resisting temptations that is ego depleting; it is also the entire gamut of tasks that draw on system 2: deliberating, choosing, inhibiting or exaggerating prepotent responses, and so on. Moreover, it is not just our ability to resist temptations that

[4] Holton has a second line of objection to the suggestion that self-control does not depend upon a discrete faculty. He points out (2003) that self-control requires *effort* and that the empirical evidence bears this out: agents who experience ego depletion give all the standard signals of physical arousal: increased blood pressure and pulse, greater skin conductance response (indicative of sweating), and so on. It should be noted that there is some evidence that deploying system 2 processes more generally is effortful, as measured in just these ways; see Naccache et al. 2005. This evidence cannot be considered conclusive, inasmuch as most (perhaps all) system 2 processes involve attentional control and therefore self-control. However, the very fact that self-control is a pervasive feature of system 2 processes might constitute evidence against the claim that self-control depends upon a discrete faculty.

is undermined by this depletion, but our ability to engage in further system 2 tasks. If we want to be able to better predict and control our own behavior, we must attend to *all* the causes and effects of ego depletion, not just those captured by the concept of weakness of the will. Indeed, even if our concern is avoiding weakness of the will – that is, with maintaining our resolutions – we need to attend to the full range of its causes.

That is not to say that the concept of weakness of the will does not refer. Agents make resolutions and all too often they find themselves acting in ways that conflict with those resolutions. We can justifiably retain the concept of weakness of the will to refer to this special case of the broader phenomenon (indeed, I have used the concept in just this manner throughout this paper). But the analogy to the Newtonian/relativity relation is not a good one: weakness of the will refers, but it does not guide us well as agents. For that reason, we can maintain the concept as a descriptive notion, but not as an explanatory or predictive one. All of us, scientists and ordinary agents alike, would do better to think of our behavior in terms of the broader phenomenon and not the narrower.

Conclusion

Weakness of will occurs because agents experience judgment-shift. But the causes of such judgment-shifts are not limited to temptations to break resolutions, and the effects of temptations are not limited to causing weakness of the will. Rather, weakness of the will is caused by the depletion of system 2 resources, throwing the agent back on to the more meager and stereotyped system 1; moreover, successfully resisting temptations to break resolutions has effects on the availability of system 2 resources. Weakness of will is simply a significant and salient manifestation of a much broader phenomenon, whereby rational agents come to act in ways that do not reflect the range and power of their rational processes.

If the forgoing is correct, weakness of the will, as a folk psychological concept, does not correspond to a psychological kind. In postulating its existence, we do not cut our cognitive nature at its joints. The effects of ego depletion are not limited to self-control nor are its causes: if we wish to be able to explain our failures of practical rationality, we should abandon the notion. Moreover, if we wish to improve our practical rationality, by increasing our propensity to act as we – in a cool hour – judge we ought, then we need to be aware of the loss of self-control as a manifestation of a broader phenomenon. Thus both our practical interests and our explanatory projects require us to abandon

the notion of weakness of the will in favour of a focus on the interrelationships between system 2 and system 1 processes.[5]

References

Ainslie, G. 2001: *Breakdown of Will*. Cambridge: Cambridge University Press.

Banfield, J., Wyland, C.L., Macrae, C.N., Munte, T.F. and Heatherton, T.F. 2005: The cognitive neuroscience of self-regulation. In R.F. Baumeister and K.D. Vohs (eds.), *The Handbook of Self-Regulation*. New York: Guilford Press.

Baumeister, R.F., Bratslavsky, E., Muraven, M. and Tice, D.M. 1998: Ego-depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74: 1252–1265.

Baumeister, R.F. 2002: ''Ego Depletion and Self-Control Failure: An Energy Model of the Self's Executive Function''. *Self and Identity* 1: 129–136.

Baumeister, R.F. and Vohs, K.D. 2007: Self-Regulation, Ego Depletion, and Motivation, *Social and Personality Psychology Compass* 1: 1–14.

Baumeister, R.F., Sparks, E.A., Stillman, T.F. and Vohs, K.D. 2008: ''Free will in consumer behavior: Self-control, ego depletion, and choice''. *Journal of Consumer Psychology* 18: 4–13.

Davidson, D. 1970: How is Weakness of the Will Possible? In Joel Feinberg (ed.), *Moral Concepts*. Oxford: Oxford University Press.

Dodd, D. 2009: ''Weakness of Will as Intention-Violation''. *European Journal of Philosophy* 17: 45–59.

FitzPatrick, W. J. 2008: ''Moral Responsibility and Normative Ignorance: Answering a New Skeptical Challenge''. *Ethics* 118: 518–613.

Gailliot, M.T. and Baumeister, R.F. 2007: ''The Physiology of Willpower: Linking Blood Glucose to Self-Control''. *Personality and Social Psychology Review* 11: 303–327.

Gailliot, M.T., Baumeister, R.F., DeWall, C.N., Maner, J.K., Plant, E.A., Tice, D.M., Brewer, L.E. and Schmeichel, B.J. 2007: ''Self-Control relies on glucose as a limited energy source: Willpower is more than a metaphor''. *Journal of Personality and Social Psychology* 92: 325–336.

Gilbert, D. 1991: ''How Mental Systems Believe''. *American Psychologist* 46: 107–119.

[5] I would like to thank Richard Holton for extremely helpful discussions of his view and of mine. This paper has also been improved immeasurably by the comments of an anonymous referee for *Philosophy and Phenomenological Research*. The author gratefully acknowledges the support of the Wellcome Trust.

Hamilton, R., Hong, J. and Chernev, A. 2007: ''Perceptual Focus Effects in Choice''. *Journal of Consumer Research* 34: 187–199.
Holton, R. 1999: ''Intention and Weakness of Will''. *Journal of Philosophy* 96: 241–262.
—— 2003: How is Strength of Will Possible? In S. Stroud and C Tappolet (eds.), *Weakness of Will and Practical Irrationality* (Oxford: Clarendon Press) pp. 39–67.
—— 2004: ''Rational Resolve''. *Philosophical Review* 113: 507–35.
Kahan, D., Polivy, J. and Herman, C.P. 2003: ''Conformity and Dietary Disinhibition: A Test of the Ego Strength Model of Self-Regulation''. *International Journal of Eating Disorders* 33: 165–171.
Masicampo, E.J. and Baumeister, R.F. 2008: ''Toward a Physiology of Dual-Process Reasoning and Judgment''. *Psychological Science* 19: 255–260.
McIntyre, A. 2006: What is Wrong with Weakness of Will? *Journal of Philosophy* 103: 284–311.
Mead, N.L., Baumeister, R.F., Gino, F., Schweitzer, M.E. and Ariely, D. 2009: ''Too Tired to Tell the Truth: Self-Control Resource Depletion and Dishonesty''. *Journal of Experimental Social Psychology*. 45: 594–597.
Mele, A. 1995: *Autonomous Agents*. Oxford: Oxford University Press.
Mischel, W., Ebbesen, E.B. and Zeiss, A.R. 1972: ''Cognitive and Attentional Mechanisms in Delay of Gratification''. *Journal of Personality and Social Psychology* 21: 204–218.
Mischel, H.N and Mischel, W. 1983: ''The development of children's knowledge of self-control strategies''. *Child Development* 54: 603–619.
Mischel, W., Shoda, Y. and Rodriguez. 1989: ''Delay of gratification in children''. *Science* 244: 933–938.
Naccache, L., Dehaene, S., Cohen, L., Habert, M.-O., Guichart-Gomez, E., Galanaud, D. and Willer, J.-C. 2005: ''Effortless control: executive attention and conscious feeling of mental effort are dissociable''. *Neuropsychologia* 43: 1318–1328.
Nickerson, R.S. 1998: ''Confirmation bias: A ubiquitous phenomenon in many guises''. *Review of General Psychology* 2: 175–220.
Rachlin, H. 2000: *The Science of Self-Control*. Cambridge, Mass.: Harvard University Press.
Schmeichel, B.J., Vohs, K.D. and Baumeister, R.F. 2003: ''Intellectual performance and ego depletion: Role of the self in logical reasoning and other information processing''. *Journal of Personality and Social Psychology* 85: 33–46.

Schmeichel, B.J, Demaree, H.A., Robinson, J.L. and Pu, J. 2006: Ego depletion by response exaggeration. *Journal of Experimental Social Psychology* 42: 95–102.

Schwitzgebel, E. 2008: "The Unreliability of Naive Introspection". *Philosophical Review* 117: 245–273.

Smith, M. 2003: Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion. In S. Stroud and C Tappolet (eds.), *Weakness of Will and Practical Irrationality*. Oxford: Clarendon Press.

Stanovich, K. 1999: *Who Is Rational? Studies of Individual Differences in Reasoning*. Mahwah: Lawrence Erlbaum Associates.

Tenenbaum, S. 2007: *Appearances of the Good*. Cambridge: Cambridge University Press.

Vohs, K. D. and Heatherton, T. F. 2000: "Self-regulatory failure: A resource-depletion approach". *Psychological Science* 11: 249–254.

Wang, J., Novemsky, N., Dhar, R. and Baumeister, R.F. Submitted: Effects of Depletion in Sequential Choices.

Watson, G. 1977: "Skepticism about Weakness of Will". *Philosophical Review* 85: 316–39.

Webb, T. L. and Sheeran, P. 2002: Can implementation intentions help to overcome ego depletion? *Journal of Experimental Psychology* 39: 279–286.

Wheeler, S.C., Briñol, P. and Hermann, A.D. 2007: "Resistance to persuasion as self-regulation: Ego depletion and its effects on attitude change processes". *Journal Of Experimental Social Psychology* 43: 150–156.