Available online at www.sciencedirect.com
ScienceDirect
Journal of Economic Psychology 27 (2006) 681–693
JOURNAL OF Economic Psychology
www.elsevier.com/locate/joep

Which side are you on? The ethics of self-command

Daniel Read *

Durham Business School, University of Durham, Mill Hill Lane, Durham DH1 3LB, UK

Received 24 March 2006; received in revised form 12 April 2006; accepted 13 April 2006
Available online 9 August 2006

Abstract

Thomas Schelling observed that when people use self-command to prevent their future selves from acting waywardly, they effectively divide themselves into two selves with conflicting desires for the same point. In the morning, for example, a dieter may not want dessert at dinnertime but change his mind when at table. This raises the question of which self is authentic or rational. Or, if we are asked to help that person to achieve his goals, which side should we take? This article considers several proposed answers to this question, and argues that they do not provide a satisfactory answer to the question of how we should or how we do take sides. It is suggested that taking sides in intrapersonal conflict is more of a judgment about the "rightness" of an action, than a consideration of what it is that the person "really" wants.

JEL classification: D63; D90

PsycINFO classification: 2340

Keywords: Hyperbolic discounting; Impulsivity; Personal identity

1. Introduction

I do not consider the injection of heroin or the smoking of nicotine to raise any issue of rationality. It is only when the user of heroin or nicotine makes a serious attempt

* Tel.: +44 191 334 5454.
E-mail address: daniel.read2@durham.ac.uk

0167-4870/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.joep.2006.04.005

> to stop and has difficulty doing so, suffering occasional relapse or suffering torment on the verge of relapse, perhaps attempting to restructure his or her environment or his or her incentives, that the issue arises whether some preferences are "true" and some are interlopers, whether fulfilling one preference is rational and fulfilling an opposing or alternating preference is not (Schelling, 1996, p. 252).

Thomas Schelling has long been concerned with the ancient problem of *intrapersonal conflicts*, and the tactics used to resolve them (Schelling, 1978, 1984a, 1984b, 1992, 1996). These are conflicts that arise over a single decision, when we contemplate it at different times. We know that we will later have a choice to make, and right now we hope that when the time comes we will choose O, but fear that we will choose X. To increase the likelihood of choosing O, we take steps to make X unavailable, or at least less attractive, to the later self who will actually have to make the choice. Examples abound. We don't want to drink tomorrow, so empty our bottles down the sink. We don't buy ice cream *because* we love it, we switch off our email, and cut up our credit cards. These are all examples of what Schelling calls *anticipatory self-command.*

The desire to deploy anticipatory self-command means that an individual can be divided into at least two selves that exist sequentially, and who differ in their preferences for what action should be taken at the same time.[1] The self that exists today wants to ride the rollercoaster tomorrow, but the self that exists tomorrow will be too afraid. The today self wants to thwart the tomorrow self, so he invites a friend to share the rollercoaster ride, knowing that tomorrow's self will be too embarrassed to back out, even though he will be mortified that today's self made that rash invitation. Because rational choice theory depends so much on such axioms as completeness and transitivity, it is natural to ask whether a single "genuine" or "correct" or "rational" preference underlies this intrapersonal difference of opinion. That is, is there some way of knowing whether this conflicted person *should* ride the rollercoaster or back out?

As demonstrated by the epigraph, this search for the genuine self is a central theme in Schelling's writings on self-command. Characteristically, Schelling does not reach any conclusions. Rather, his method is to describe a multitude of cases, each offering a different challenge to possible ways of defining this self. He does, however, provide us with a goal: "If somebody now wants our help later in constraining his later behaviour against his own wishes at that later time, how do we decide which side we are on?" (Schelling, 1984a, p. 87). In this article I examine several possible decision rules or criteria that might help us make this decision.

2. The framework

Throughout this article, I will refer to three "agents" as depicted in Table 1. These agents are defined in terms of a division of life into intervals, over each of which the person has preferences for actions taken at a specific time, denoted t_A. The *acting-agent* is the one who will or won't perform a target action, such as riding the rollercoaster, smoking the cigarette, or eating the dessert. The *pre-agent* is one who anticipates the possible actions

[1] It is this fact that has led economists to use the language of multiple selves (e.g., Strotz, 1956), and even to discuss the problem in game theoretic terms. Schelling's (1960) first published mention of intrapersonal conflict is in reference to Luce and Raiffa's (1957) *Games and Decisions.*

Table 1
Preference patterns defined over three agents representing a person's preferences for actions taken at different times, with preferred (O) and problematic options (X)

Preference pattern	Agent/time			Judgment
	Pre-agent (t_{PRE})	Acting-agent (t_A)	Post-agent (t_{POST})	
Consistent	O	O	O	"I did the right thing"
Local-conflict[a]	O	X	O	"I wish I had not done that"
Melancholy[a]	O	O	X	"I wish I had"
Liberation[a]	O	X	X	"I am glad I took the wrong turn"

[a] Intrapersonal conflict.

that might be taken, has a preference over those actions, and will be the one to deploy anticipatory self-command, should it be judged necessary. The *post-agent* is one who reflects back on what has taken place, and may even suffer or enjoy some of the consequences.

Table 1 shows four preference patterns over the three agents, including three kinds of conflict, which arise when the preferences for t_A do not coincide. The letter O represents the option preferred by the pre-agent, while X represents the "problematic" option. Consistent preferences are the uncontroversial ones – all agents agree that the best choice at t_A is O. This is the normal situation. You hope to take a holiday on June 3, you do take the holiday, and you are pleased that you did. The pattern designated as "local-conflict" is the one that has been the focus of most discussion, with the acting-agent disagreeing with both the pre- and post-agent about the best choice at t_A. Although the pre-agent hopes to not drink heavily tonight, the acting-agent does drink, and the post-agent regrets it. Two other forms of conflict, to which we will return, are also depicted in Table 1.

The question of "taking sides" comes down to whether the preferences held by a single agent or a coalition of agents should take precedence. Is there a genuine, or legitimate, or representative, or right-thinking self that can be identified by examining the conflicting preferences of the person? Many researchers have proposed rules for identifying this self. Some approaches view the characteristics of the option as being the deciding factor, with some options being better than others and therefore the agent that prefers the wrong option should either be disregarded (if possible) or managed. Others propose that this identification can be made on the basis of the *relationships* between agents, with some choice patterns or preference relationships pointing out which side is the genuine article. In this article, I will consider both kinds of solution.

3. Consequentialism

> A thing is said to promote the interest, or to be for the interest, of an individual, when it tends to add to the sum total of his pleasures; or, what comes to the same thing, to diminish the sum total of his pains (Bentham, 1789/1987, p. 66).

Most people will agree that it would be better for an alcoholic to not drink. Drinking is pleasurable while it lasts, but its after-effects might ruin his life. One reason to take sides with the non-drinking self, therefore, is that the *consequences* of not drinking are better than those from drinking. These consequences are not just those experienced at t_A but those experienced at all other times as well. Although it is clear – as shown by the passage cited at the beginning of the article – that Schelling does not hold that this alone is a sufficient

basis for taking sides, I suggest that it is the foundation of our untutored judgements of what side to take, and is therefore a good place to begin.

The passage from Bentham describes the "purest" form of consequentialism. The criterion is to maximize total pleasure minus total pain. Imagine someone who will soon face a choice between drinking or not-drinking a double-whiskey. In his current role as pre-agent, he wants to resist the temptation to drink, and can draw on the tactics of self-command to do so. If he chooses the whiskey, the momentary pleasure or pain that he will experience at each instant during his life will *partly* depend on this choice, as well as on every other choice he makes. Two possible scenarios are as follows: If he is going to drink, this makes him less happy than he would otherwise be in the period leading up to t_A, at t_A he will be pretty jolly, and then afterwards his pleasure will be tainted with some regret; On the other hand, if he is not going to drink, he will be slightly happier leading up to t_A, less happy at t_A, and perhaps feel some relief afterwards. Whether it is better to drink or not depends on whether the life lived with the whiskey drinking is better than the life lived without it.[2] This, in turn, is established by summing up all the moments of pleasure and pain in both lives. The life with the greatest total is the one he should choose.

Consequentialism does not need to focus on pleasure. For instance, although Mill was a disciple of Bentham's, he famously argued for "higher" and "lower" kinds of pleasure, with the higher pleasure being more important than the lower (Mill, 1863/1987). Likewise, we might believe that certain feelings, such as those of self-respect or self-actualization, are more important than pleasure itself. But regardless of such details, the consequentialist approach is based on the general view that the best options are those that maximise a (possibly weighted) sum of good effects minus bad ones.

As described, consequentialism does not offer a practical solution to the problem of taking sides. While it gives a (controversial) definition of what option is best, it does not give us a recipe for establishing which side in an intrapersonal conflict is best according to that definition. One of Schelling's favourite examples is of Captain Ahab (from the 1930 film, not from Melville) whose crew hold him down while they cauterize the leg he lost to Moby Dick. Ahab orders them to stop, but they perform the operation anyway. Reflecting on this, Schelling asks, "How do we know if an hour of extreme pain is more than life is worth?" The answer, of course, is that we do not. The consequentialist approach is an idealisation that specifies which side to take if we had information that we cannot obtain. To make the decision, we have to fall back on the same intuitions that we set out to justify.

We can illustrate the difficulty with a case of addiction. William Burroughs spent much of his life struggling with heroin addiction (Morgan, 1988), yet he became one of the most influential writers of his time. Years later, he had this to say: "I have never regretted my experience with drugs. I think I am in better health now as a result of using junk at intervals than I would be if I had never been an addict. When you stop growing you start dying. An addict never stops growing" (Burroughs, 1953/1977, p. 13).

Burroughs claimed that the lifetime benefit from heroin use exceeded that from non-heroin use. Perhaps he was wrong, but without knowing what would have happened to him without the addiction, we cannot definitively side with either his junky or non-junky selves by drawing on consequentialist reasoning. The problem is that this decision rule is based

[2] Actions also affect the probability of taking other actions – one drink leads to another. These effects will also be included in the total life analysis just discussed.

on the *total* consequences of an act. Other rules for taking sides draw on local evidence – what preferences we have and when, what actions we take and when, and the attitudes we have toward those preferences and actions.

4. Hyperbolic discounting

> To-day it will be rational for a man to jettison his "optimal" plan of yesterday, not because his tastes have changed in any unexpected way nor because his knowledge of the future is different, but because to-day he is a different person with a new discount function – the old one shifted forward in time (Strotz, 1956, p. 173).

Strotz (1956) is describing hyperbolic discounting, according to which acting-agents discount the future by overweighting the benefits that will be received at t_A, and underweighting those that will be received later. When an action is some distance away, we make a consumption plan that gives reasonable weight to all points in the future. But when the action is imminent, we revise the plan to bring consumption forward. That is, we take more pleasure and less pain now, in exchange for less pleasure and more pain in the future. A sophisticated pre-agent will disapprove and so deploy self-command to ensure that the original consumption plan is maintained.

The hyperbolic discounting account gives a rationale for siding with the pre-agent. It holds that under normal circumstances the acting-agent irrationally overweights immediate over delayed concerns, so that in cases of conflict the "correct" choice is always the option preferred by the pre-agent. According to Ainslie (1975), the conflict between agents arises when there is a choice between smaller-sooner and larger-later rewards. The struggling alcoholic who wants to be sober is choosing between the smaller-sooner rewards of intoxication, and the larger-later rewards of a fulfilled life. The pre-agent will often choose the larger-later reward, while the acting-agent can be overcome by a desire for the smaller-sooner one, a preference that Ainslie refers to as "self-defeating", and "an elementary trait that makes us irrational, at least for living in the developed world" (Ainslie, 2002, p. 3). The hyperbolic discounting account is a form of consequentialism, because it tells us when the total utility from one option is greater than that of another. By definition, the option chosen by the acting-agent can never bring more total utility than that chosen by the pre-agent, and will often bring less. Therefore, in cases of conflict we should side with the pre-agent.

This decision rule fails because the premise that all intrapersonal conflict is due to hyperbolic discounting is wrong.[3] There is at least one other source of conflict that often acts in the opposite direction. The information available to the acting-agent about the local consequences of a specific choice will often be better than the information available to the pre-agent. When a dieter changes his mind and has tiramisu after promising not to, it might be because he is weak-willed, or it might be because he has only now realized how appealing the tiramisu is. Schelling (1984a) discusses precisely this problem when describing the

[3] This is not to say that Ainslie or any author takes precisely this position, although they often appear to approach very close to it (e.g., Frank, 1988). The point is that for the theory of hyperbolic discounting to allow us to take sides, we either need to say that pre-choice is always best, or else to be able to say when hyperbolic discounting applies and when it does not. This, of course, only shifts the problem of taking sides to the problem of deciding when hyperbolic discounting applies.

conflicting preferences of someone who plans to go running in the morning, but backs out when the time comes: "It is not clear whether the straight fellow who resolves to run three miles before breakfast enjoys such a far horizon that he can appreciate the benefits of elderly good health, or merely has such a short memory that he forgets how disagreeable its going to be, every morning in perpetuity, to spend thirty minutes gasping for breath" (Schelling, 1984a, p. 63). There is a strong information asymmetry here, and the only concrete and certain information concerns today's pain – the benefits from running are anticipated and uncertain – so it is difficult to justify taking sides with the ambitious pre-agent's preferences.

This is a specific case of a general problem common to many proposed mechanisms for taking sides. It holds that pre-agents will generally make the "right" decision, even when they will not know the situation on the ground when they must be enacted. The acting-agent's better information about local circumstances is disregarded. Notice how similar this is to the problem of centralized economic planning as discussed by Hayek (1945). Who is better placed to judge what is best than the acting-agent who knows how it will feel to put a well-intentioned plan into practice? The pre-agent does not have this information, and might therefore be tempted to put disproportionate weight on its *own* preferences, and attempt to impose them on the acting-agent.

Naturally, there will be *some* cases in which disagreement between the pre-agent and acting-agent are due to the excessive discounting or myopia of the acting-agent. But not every case is like this, and so even if we believe that myopia on the part of the acting-agent is a good reason to side with the pre-agent, it may be impossible to know when myopia is the cause, and when something else is.

5. Time spent in one state of preference or another

> The time interval B [t_A in our notation] is not the appropriate benchmark for deciding what the person ought to do because B is not a representative sample of her view of the matter. … I suggest that, often, what makes resisting the temptation and taking the larger reward the preferred option is the person's preference for a majority of the time; it is her (reasonably) stable preference. The other is her preference at a non-representative moment (Nozick, 1993, p. 16).

The hyperbolic discounting and total utility perspectives both hold that some objective feature of the choice object is the basis for judging whether a choice is rational or not. The hyperbolic discounting approach takes it as axiomatic that the acting-agent is less concerned with our total good than is the pre-agent. The consequentialist approach does not make such an assumption, but holds that a separate measurement can tell us whether or not the correct option was chosen. Nozick's (1993) approach is to identify the correct choice option with an identifiable feature of the person's overall attitude towards that option. That feature is that it is a "representative" or stable preference, meaning a preference held for a majority of the time. We should take sides with this stable preference.

Holding a majority preference for O does not necessarily mean that O is *ever* chosen. Consider a struggling alcoholic who mostly prefers *not to* have his 10:00 a.m. eye-opener. The evening before, while finishing-off a bottle of rye, he vows that tomorrow morning he will turn over a new leaf and won't have the first drink. In the morning he predictably changes his mind. Yet later, as he continues drinking, he "wishes" he had stuck to his

resolution. In Nozick's account, that person's stable preference was to *not* have the eye-opener, because the majority of moments would vote for not drinking at 10:00 a.m. Generally, the concept of majority preference is defined over single acts performed at specific moments, and not over classes of acts. Right now, our fanciful alcoholic wants a drink, yet dreads every drink before it is drunk, and regrets it afterward.

The principle that the representative self is defined by majority preference applies to many more cases than those intrapersonal conflicts discussed by Schelling. There will be many times when an option will draw support for a majority of the time, yet not obviously be the option we will consider the best. Consider the cases of Melancholy and Liberation from Table 1, in which the choices dictated by the pre-agent are regretted, or where the changed preferences of the acting-agent persist. These cases are exemplified by the literary staple of a man devoted to his work who does not plan to fall in love but who is unexpectedly confronted with this heretofore resisted opportunity. All three inconsistent patterns from Table 1 are possible. He might succumb to temptation, and afterwards wish he had not; he might succumb and be glad; or he might resist and regret it (cinematic examples are, respectively, *The Blue Angel*, *Pickpocket*, and *Remains of the Day*). In the latter two cases, the choice pattern itself does not tell us whether the representative or longest-lasting self is the one choosing *O* or *X*. It depends, indeed, on how much of one's life is spent before t_A and how much afterwards. If our hero is young, then the *X* preference holds sway for most of his life so this is the majority preference, while this will change as he ages.

I believe it is unlikely that in this situation, or others like it, we will takes sides based on numerosity.[4] Indeed, Nozick himself will not want to. His own account is tentative (he notes that "this criterion is meant to be defeasible [subject to refutation by new evidence], not conclusive"), and it is carried out as part of a discussion of more traditional local-conflict preference reversals. But this suggests the total-time rule does no theoretical work. When the total-time rule is coupled with a local-conflict preference reversal, it indicates that the acting-agent is making the wrong decision. But if the total-time rule is coupled any other type of preference inconsistency, it does not.

6. Asymmetry of anticipatory self-command

> I do not know of any important cases in which the following criterion is inapplicable: the authentic self is the one who is capable of acting strategically towards the other self or selves. In other words, I deny that there are real-life examples of two selves – the self who wants to stay drunk, and the self who wants to stay sober – engaging in mutual strategic interaction (Elster, 1985, p. 92).

Elster (2000) offers the criterion of "authenticity", one of many terms also used by Schelling to label the self with whom we should take sides. The idea is that there is an authentic self that can be identified by some function of intertemporal preferences, and with whom we should side. Elster's empirical claim, as it is given above, appears correct – I cannot think of counterexamples, nor have other writers offered any. This does not mean

[4] The Nozick criterion also has trouble with the William Burrough's case, which involves a preference pattern such that OXOX, where a late-appearing agent is glad of former dissolution, even if earlier agents were less happy and struggled to resist it.

that the acting-agent might not often wish he could go back and change the commitments of the pre-agent, but that the acting-agent usually does not commit future selves to choose the same things he wants. An alcoholic might curse himself for pouring liquor down the sink yesterday, but once he buys more liquor he will not try to prevent his future self from pouring it down the toilet.[5]

When discussing self-command (he called it "self-control"), Skinner (1953) explained how this asymmetry arises. For him, self-command is used when behaviour has positive consequences for a person at one point in time, but aversive consequences at another. For example, a person might be rewarded by the pleasures of drinking on Saturday night, but during the week feel shame at his behaviour and try to prevent it from being repeated on the next Saturday. There is therefore a conflict-of-interest between two agents, one who is reinforced by his own behaviour and seeks to perpetuate it, and another who is punished by the same behaviour and seeks to end it. Because the first agent is not punished by the second, he makes no attempt to change it.

Despite the plausibility of the empirical claim, strategic asymmetry does not satisfactorily define the class of authentic choices. One reason is that (like the total time criterion) it neglects the role of actions in defining the authentic self, and therefore can identify it with a preference that is never made manifest. We can illustrate this with the fictional case of Garcin from Sartre's *Huis Clos*. Garcin (now in hell) reflects on his life, when he deserted from the army and had been sentenced to die for his desertion. He rationalised his desertion as being the courageous act of a conscientious objector, but was worried that it might have been an act of cowardice. He tells his story to one of his fellow damned:

> GARCIN: … always I harked back to the one thing certain – that I had acted as I did, I'd taken that train to the frontier. But why? Why? Finally I thought: My death will settle it. If I face death courageously, I'll prove I am no coward.
>
> INEZ: And how did you face death?
>
> GARCIN: Miserably. Rottenly. Oh, it was only a physical lapse – that might happen to anyone; I'm not ashamed of it. Only everything's been left in suspense forever (Sartre, 1944/2000, p. 92).

Garcin *desired* to be courageous and even *planned* to be, but he failed when he had to act. Sartre's view, which I suspect most of us would share, is that Garcin is authentically a coward, regardless of the plans of the pre-agent or regrets of the post-agent. We do not have to turn to such exotic examples. Smokers who want to quit can spend many years having a last cigarette – always planning to have no more, and always regretting the last one while lighting up the next (Klein, 1993). These people are no more authentically non-smokers, than perpetual dieters are authentically thin.[6]

This definition of authenticity also "stacks the deck" against treating present-oriented actions as authentic. Consider Ryle's (1954) account of pleasure. His view was that the

[5] Addicts do attempt to thwart the well-intentioned interventions of others, such as by keeping a hidden cache of supplies, but they don't attempt to thwart themselves. However, we can't rule out the possibility that this is because it is difficult to hide something from your future self. In other words, addicts in their addictive state may have the desire, but not the means or the wherewithal, to deploy self-command to thwart themselves from deploying self-command in the future.

[6] I am reminded of a friend who referred to herself as a "non-practicing vegetarian".

Table 2
Second-order preferences of the pre-agent and the acting-agent do not differ, but their first-order preferences differ

Preference	Pre-agent (t_{PRE})	Acting-agent (t_A)
Second-order preference	"no-drug $\succ$ drug" $\succ$ "drug $\succ$ no-drug"	"no-drug $\succ$ drug" $\succ$ "drug $\succ$ no-drug"
First-order preference for t_A	no-drug $\succ$ drug	drug $\succ$ no-drug

greatest pleasure occurs when we are fully engaged in an activity, and so absorbed that we neither want to do anything else, nor entertain the possibility of doing so. Most will agree that a life filled with such periods is likely to be a good one. Yet, it is precisely when so absorbed that we will not spend our time attempting to constrain the actions of our future selves.[7] The asymmetric self-command view is biased against these present-oriented states, because commitments are always made *for* these states, when in other less-pleasurable ones. To take sides with the pre-agent, we have to trust it to know when self-command is appropriate and when it is not.

7. Second-order preferences

> The unwilling addict identifies himself, through the formation of a second-order volition, with one rather than with the other of his conflicting first-order desires. He makes one of them more truly his own and, in so doing, he withdraws himself from the other (Frankfurt, 1971, p. 13).

Frankfurt (1971) used the term "second-order preference" to describe preferences over preference orderings, rather than, as in the case of first-order preferences, over choice options. In the passage above, he introduces the familiar two sides in intrapersonal conflict, and defines them in terms of differences in their first- and second-order preferences. Given a choice between taking or not taking a drug, the addict prefers to take the drug. This is his first-order preference. Given a choice, however, he would rather have the preference ordering "no drug $\succ$ drug" than "drug $\succ$ no drug". This is his second-order preference. A second-order preference is not merely a manifestation of "mixed feelings", but a preference ranking such that if this person could choose from a menu of preferences he would choose to prefer not-drinking over drinking.

Although Frankfurt did not discuss second-order preferences in the context of self-command, George (1998) has done so in a commentary on Schelling (1996). George argues there are many situations in which, while the first-order preferences of the pre-agent and the acting-agent differ, their second-order preferences do not. The idea is illustrated in Table 2. The second-order preference for both the pre-agent and the acting-agent is to prefer to not take drugs, but the first-order preference (for consumption at t_A) differs in the usual way.

George (1998) considers under what circumstances we can side with the pre-agent over the acting-agent. His view is that we can take sides with the pre-agent if his use of self-command will change the anticipated *first-order* preferences of the acting-agent. One way this

[7] We might make a commitment, such as to marriage, but not because, at the moment of choice, we are fearful that we will later change our mind, but because we are certain we will not.

can happen is through the process of "out of sight, out of mind". By hiding drugs or making them unavailable or avoiding contexts in which they are usually taken, the unwilling addict may simply not want them at t_A. Regardless of whether this is true for drug addicts, it is undoubtedly true for many everyday examples of procrastination. For instance, I am currently sitting at a computer without an Internet connection. I feel no desire to "check my mail" or "read the news". Yet if I had an Internet connection I would feel those desires, and would succumb. The fact that the Internet is unavailable means that my current first-order preference is for no-Internet, yet if it were available my first-order preference would be for Internet. George's view is that in cases like these we should side with the pre-agent who deploys self-command. I believe this is uncontroversial. But it is uncontroversial because, as in my case, there is no conflict – both agents want to do the "right" thing, and the agent who wants to surf the net is counterfactual.

George's analysis therefore leaves unresolved more difficult cases such as Frankfurt's unwilling addict, in which both the pre-agent and acting-agent have the same second-order preferences, even if their first-order preferences differ. At t_P he wants to not take drugs at t_A, but at t_A he wants to take drugs. But, if we have correctly characterised him as someone who wants to quit, then at both t_P and t_A he would choose to have the preference "no-drug $\succ$ drug". Can we take sides with this second-order preference?

I think that if such cases exist we are likely to agree we can. It would be difficult to deny sympathy with the second-order preference of a drug addict who, while reaching for his works, piteously says, "I wish I could beat this craving". We might also side with the inauthentic Garcin if he has a second-order desire to be courageous.

This solution faces two difficulties, however, which revolve around the question of whether this characterisation of conflicting first- and second-order preferences is correct. The first point is closely related to one made earlier. When fully engaged in an activity we are usually not reflecting on whether our preferences are the ones we wish to have. Consider my own example of Internet overuse. My pre-agent may try to prevent me from using the internet, and will wish that I preferred "no-internet $\succ$ internet" to "internet $\succ$ no-internet". But *while* using the internet I never have these second-order preferences. Frankfurt (1971) uses the term "wanton" to describe a being incapable of second-order preferences, and we might say that when we are engaged in most activities (rather than anticipating them or reflecting on them) we are momentarily wanton. That is, the acting-agent does not *have* any second-order preferences.

A second problem remains, however, even if we assume that the acting-agent does have conflicting second-order and first-order preferences. Imagine we interrupted an addict's Internet or drug use and offered her the opportunity to change her second-order preferences. What will the addict choose? It will depend, I suggest, on the scope of the preference change. One possibility (and I think this is what we will naturally imagine the change means) is that by changing our second-order preference she will never again prefer drugs to no-drugs. But agreeing to this change would not prove that the acting-agent is dissatisfied with his current preferences. Recall that at t_A the person is also a pre-agent for all future agents, and if a decision has implications for these future agents, the person may be making a policy decision, involving a trade-off between desires for consumption at t_A, and desires for consumption at all other times. This is a common view of what self-command entails (e.g., Ainslie, 1975; Nozick, 1993; Rachlin, 1995).

To ensure that an apparent desire for a different first-order preference by the acting-agent is not, in fact, a desire by the pre-agent to control future first-order preferences, we

must assess the desire for changed preferences at t_A when they are independent of any future preferences. There are two choices to consider. The first is between:

A1: "no-drug ≻ drug" at t_A, with all future preferences unchanged;
A2: "drug ≻ no-drug" at t_A, with all future preferences unchanged.

In other words, a choice between the two preference orderings for t_A, but when the choice will have *no* implications for future preferences or choices. Specifically, any future struggle with drugs will occur just as it would have regardless of whether A1 or A2 is chosen. The second choice is between two policy decisions:

B1: "no-drug ≻ drug" at t_A and on all future occasions;
B2: "drug ≻ no-drug" at t_A, and "no-drug ≻ drug" on all future occasions.

If the addict holds the preference pattern in Table 2 she will choose A1 and B1. That is, she will change her first-order preferences at t_A *regardless* of the implications of that choice for future selves. It is an empirical question whether she will, but my prediction is that the most common choice pattern will be A2 and B2. The desire for changed preferences is the same kind of desire Saint Augustine had when he asked "God, give me chastity and continence – but not just now".

8. Discussion

> Though so profound a double-dealer, I was in no sense a hypocrite; both sides of me were in dead earnest; I was no more myself when I laid aside restraint and plunged in shame, than when I laboured, in the eye of day, at the furtherance of knowledge or the relief of sorrow or suffering (Stevenson, 1886/1992, p. 83).

> Sometimes, but not always, it is easy know which is Jekyll and which is Hyde (Schelling, 1984a, p. 61).

This article has investigated the question of how to take sides in intrapersonal conflict. I have argued that several possible solutions are unsatisfactory on two counts. First, they do not reflect our judgments about which side to take – in every case it seems that we have to fall back on intuitions separate from the principles to make the decision. Second, despite this philosophical indeterminacy, people generally have little difficulty taking sides. Although Schelling (1984a) says it is only "sometimes" easy to know which is Jekyll and which is Hyde, his articles contain very few truly difficult cases even though he is actively *seeking* those cases. I will conclude, therefore, by considering why we find taking sides so easy.

My suggestion is that we have prior beliefs, based on a number of criteria,[8] about what is "best" for someone, and that we side with the agent who chooses these best acts. Moreover, this is true even when there is no evidence that the alternative act has or will cause harm. For instance, we will want to help a smoker to quit, even if he has formerly been per-

[8] The criteria include judgments about the person's welfare, but will also include ethical and moral judgments, preferences for social convention, and the self-interest of person making the judgment. One reviewer also mentioned that envy might lead us to disapprove of Hyde, since we might feel he is having more fun than us.

fectly happy smoking and displays no negative effects from smoking. Yet if a non-smoker wanted to become a smoker, but was unable to acquire the habit because smoking made him nauseous and needed someone to help him through those awkward moments, we would be much more reluctant. Similarly, we will usually take the side of the dieter when he wants to diet, of the alcoholic when she is struggling to get off the bottle, and of the patient who will feel better *after* the painful surgery. In every case, we choose sides based on our prior judgments about what is best for the person.

This suggestion, however, raises a problem that is reflected in the juxtaposition of passages that opened this section. Schelling proposes that we want to decide "which is Jekyll and which is Hyde". But as the passage from Stevenson shows, neither Hyde nor Jekyll (except in the final throes of duality where the drug ceased to work) were unsatisfied with their identity: there was no dual self created by a struggle for self-command. But this does not mean that we won't side with Jekyll over Hyde, even though he does not ask us to.

Indeed, taking sides does not even require *sides*. We would side with Dr. Jekyll even if there were *only* a Mr Hyde. Imagine an alcoholic friend asks you to lend her money for alcohol, or to go to the store and buy her a bottle, or even to share a drink. Whether our friend has previously expressed a desire not to drink on this occasion will not be a major determinant of our sympathies. If we believe she is drinking herself into an early grave, we will side with a self who desires not to drink, even when that self is entirely counterfactual. That is, we want people to do what is right, period.

But what of the difficult cases? Some of the ones given by Schelling are not difficult for us, but only for theory. While we may not be able to state with certainty that Captain Ahab will get greater benefit from having his wound cauterized than from dying of gangrene, nobody (not even Schelling) will side with the self who wants the infection. Other cases are difficult because we are indifferent, because we have made no prior evaluations, or because they involve an unresolvable conflict of values. But the problem is not in the fact that there is an intrapersonal conflict, but only in that we do not know what is the right thing to do even in the absence of conflict.

This is not to say that we are not influenced by intrapersonal conflict, but the influence is more in the form of giving us information rather than defining the different sides with which we can sympathise. The conflict plays two informational roles: first, in ambiguous situations it signals the presence of a problem and thus that there is a question of taking sides; second, it tells us there is an active desire to "help oneself". The first role is revealed when a person that we don't think is in any way overweight shows that he is struggling to overcome the local need for forbidden food. Because we don't recognise he has a problem, it is only in this way that we learn there is one (and notice how half-hearted our response is likely to be to his plight). The second role is seen in the differential response to someone in clear need of help who pleads for that help compared to someone who does not. The plea tells us that our help is more likely to bear fruit.

References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin, 82*, 463–496.

Ainslie, G. (2002). The effect of hyperbolic discounting on personal choices. Paper presented at annual convention of The American Psychological Association, August 22, 2002.

Bentham, J. (1789/1987). An introduction to the principles of morals and legislation. In J. S. Mill & J. Bentham (Eds.), *Utilitarianism and other essays* (p. 789). Harmondsworth: Penguin (original work published 1789).

Burroughs, W. S. (1953/1977). *Junky*. Harmondsworth, UK: Penguin (original work published 1953).
Elster, J. (1985). Review of T. C. Schelling, Choice and Consequence. *Journal of Economic Behavior and Organization, 6*, 90–92.
Elster, J. (2000). *Ulysses unbound*. Cambridge: Cambridge University Press.
Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. New York: W.W. Norton.
Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*, 5–20.
George, D. (1998). Coping rationally with unpreferred preferences. *Eastern Economic Journal, 24*, 181–194.
Hayek, F. (1945). The use of knowledge in society. *American Economic Review, 35*, 519–530.
Klein, R. (1993). *Cigarettes are sublime*. Durham, NC: Duke University Press.
Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley.
Mill, J. S. (1863/1987). Utilitarianism. In J. S. Mill & J. Bentham (Eds.), *Utilitarianism and other essays*. Harmondsworth: Penguin (original work published 1863).
Morgan, T. (1988). *Literary outlaw: The life and times of William S. Burroughs*. New York: Henry Holt.
Nozick, R. (1993). *The nature of rationality*. Princeton, NJ: Princeton University Press.
Rachlin, H. (1995). Self-control: Beyond commitment. *Behavioral and Brain Sciences, 18*, 109–159.
Ryle, G. (1954). Pleasure. *Proceedings of the Aristotelian Society*(Supplement 28), 135–146.
Sartre, J. P. (1944/2000). *Huis Clos and three other plays*. Harmondsworth: Penguin (original work published 1944).
Schelling, T. C. (1960). The strategy of conflict.
Schelling, T. C. (1978). Egonomics, or the art of self-management. *American Economic Review: Papers and Proceedings, 68*, 290–294.
Schelling, T. C. (1984a). *Choice and consequence*. Cambridge, MA: Harvard University Press.
Schelling, T. C. (1984b). Self-command in practice, in policy, and in a theory of rational choice. *American Economic Review, Papers and Proceedings, 74*, 1–11.
Schelling, T. C. (1992). Self-command: A new discipline. In G. Loewenstein & J. Elster (Eds.), *Choice over time* (pp. 167–176). New York: Russell Sage Press.
Schelling, T. C. (1996). Coping rationally with lapses from rationality. *Eastern Economic Journal, 22*, 251–269.
Skinner, B. F. (1953). *Science and human behavior*. New York: Free Press.
Stevenson, R. L. (1992). The strange case of Dr. Jekyll and Mr. Hyde. Retrieved March 16, 2006. Available from http://www.gutenberg.org/dirs/etext92/hyde10.txt (original work published 1886).
Strotz, R. H. (1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies, 23*, 165–180.