Draft of April 3, 2012.

Response to Rabin.[1]

Alex Voorhoeve
London School of Economics

In his contribution to this volume, Matthew Rabin offers an insightful analysis of three biases:

> 1. *projection bias*: the tendency to judge one's future wants by one's current desires;
> 2. *present bias*: the tendency to do what yields an attractive outcome in the moment in a way that is contrary to one's long-term interests; and
> 3. *naïveté about present bias*: the failure to recognise one's susceptibility to present bias.

Rabin argues that these biases are irrational and details the self-harm they cause in the presence of habit-forming choices. (This harm is understood in a liberal manner as a lesser degree of satisfaction of an individual's informed, long-term preferences.) He also argues that, to prevent this harm, governments should tax behaviour that leads to unhealthy habits.

Rabin's proposal invites the charge of paternalism. This charge is unwelcome to governments. To avoid it, they commonly appeal not to self-harm, but to the costs to *others* of unhealthy behaviour (e.g., in terms of higher state-funded health care provision) to justify intervention in choices regarding smoking, exercise, and the consumption of fatty and sugary foods (de Marneffe, 2006).

I believe it is indeed important to avoid paternalistic interference when such interference limits informed, rational individuals' autonomy. However, with regards to unhealthy behaviour, the strategy of avoiding paternalism by appealing to harms to others has two weaknesses. First, it is unclear whether commonly targeted forms of unhealthy behaviour really do impose net costs on others. For example, a recent study concludes that while smoking and obesity reduce the life expectancy of 20-year-olds by 8 and 5 years respectively and increase health spending related to these behaviours, "*total* lifetime health spending was greatest for healthy-living people, lowest for smokers, and intermediate for the obese" (van Baal *et al.* 2008, p. 249, emphasis added).

Second, the focus on third-party effects ignores the ways that interventions to prevent self-harm can be justified consistently with respect for autonomy. One way is to seek the consent of those who are interfered with. After all, an intervention to which an individual consents does not represent paternalistic interference (Feinberg 1986, chapter 17). If Rabin is right about people's tendencies to choose outcomes that are contrary to their own considered preferences and about the ability of taxes to counteract these tendencies, then an enlightened majority might well consent to such taxes (the dissenting minority would then not be taxed *paternalistically*, but rather *for the sake of others*, who wish the taxes on themselves).

But even when such consent cannot be secured, one may appeal to so-called *soft paternalism*, which holds that the state has reason to constrain self-harming conduct without the consent of the people it constrains "*when but only when* that conduct is substantially non-voluntary" (Feinberg 1986, p. 12, emphasis in original). Conduct is substantially non-voluntary, in the sense here intended, when it is insufficiently informed or performed by someone insufficiently capable of rational self-governance.

Soft paternalism attractively combines a concern with promoting individual well-being with respect for the rights of individuals to govern themselves when their decisions are

[1] I am grateful to Luc Bovens, Judy Jaffe, Joe Mazor, Harald Schmidt and audiences at the LSE for helpful comments.

substantially voluntary (ibid, pp. 12-16). In this note, I shall therefore ask whether there is a soft-paternalistic justification for the taxes that Rabin advocates. My answer will be nuanced. Rabin's description of these biases as "irrational" suggests that the choices they prompt are substantially non-voluntary. However, I shall argue that Rabin's description of these biases as "irrational" is not always appropriate—sometimes, for example, they are merely a form of preference change. When they are due to the latter, the behaviour is fully voluntary, and there exists no soft-paternalistic justification for coercive intervention. I shall also argue that even when these biases *do* lead to substantially non-voluntary choices, we should prefer policies that improve self-knowledge and self-control to taxes. However, I shall note that Rabin's analysis reveals circumstances under which these autonomy-enhancing strategies will not be effective. In such cases, I shall conclude, Rabin's taxes have a soft paternalistic justification.

1. Habits

Rabin defines an activity as *habit forming* if and only if the marginal expected[2] utility a person gets at a time from consuming the good at that time is higher when she has consumed more of it in the past (p. 4 in ms.) A habit is *bad* if and only if past and current consumption has the additional effect of lowering future expected utility levels. A habit is *good* if and only if past and current consumption has the additional effect of raising future expected utility levels. Here are two examples, also represented in Table 1:

> Bad habit: smoking today increases the marginal utility of smoking tomorrow, while lowering the utility levels of smoking and not-smoking tomorrow.
>
> Good habit: running today increases the marginal utility of running tomorrow by raising the utility levels associated with running tomorrow and the utility level of not-running tomorrow.

Table 1: Utility levels and marginal utility in each period for two habits

A bad habit			
Time Action(s)	t1	t2	t3
Never smoke	6 MU=4	6	6
Always smoke	10	7 MU=5	0
Smoke at t1, give up t2		2	MU=6
Smoke at t1 & t2, give up t3			-6
A good habit			
Never run	6 MU=-3	6	6
Always run	3	7 MU=0	9
Run at t1, give up t2		7	MU=1
Run at t1 & t2, give up t3			8

2. Projection bias

Consider first a fully-informed decision-maker who, at time t1, evaluates an action by summing all present and future utilities generated by that action. In our examples, such a decision-maker would avoid the bad and choose the good habit.

[2] In this volume, Rabin does not extend his definitions to the risky context, but this seems a natural extension. I make use of this extended definition when discussing how naïveté can be advantageous in Section 3.

Now introduce projection bias—a tendency to mispredict future utilities associated with an activity by adjusting them in the direction of the current utility of that activity. An example is provided in Table 2, which focuses on the choice between always abstaining from and always engaging in the habits described above. We can see that this would result in the decision-maker incorrectly judging that indulging in the bad habit is preferable to not indulging and that not cultivating the good habit is preferable to cultivating it.

Table 2: Utility levels and marginal utility in each period as perceived by a decision-maker with projection bias at t1.

A bad habit			
Time Action(s)	t1	t2	t3
Never smoke	6	6	6
Always smoke	10	8 (real utility =7)	5 (real utility =0)
A good habit			
Never run	6	6	6
Always run	3	5 (real utility =7)	6 (real utility =9)

Rabin describes projection bias as "irrational" and believes we should counteract its effects by taxing bad habits (to reduce their perceived utility) and subsidizing good ones (to increase their attractiveness). Are these judgments justified?

One cause of projection bias may be mere lack of information. If a person lacks evidence for how something will affect her well-being, it may be sensible for her to extrapolate from her current experience of that thing. Moreover, in such cases, educating her about the consequences of her choices is better than taxing her, because the former does not limit her liberty.

However, Rabin offers evidence that not all forms of projection bias are properly ascribed to a lack of information. He cites studies which show that people's current affective state influences their judgment of what would be best for them at a future time in which they will *not* be in this state, *even when these people possess abundant evidence of their shifts in tastes*.

It is unclear what the cause is of this failure to take into account readily available information on one's future tastes. One hypothesis is that projection bias is the result of decision-making by the so-called "intuitive system" (Kahneman 2002). Such decisions are typically quick, automatic, and require little mental effort; they are also strongly influenced by current stimuli. (In this, they contrast with decisions made by the reasoning system, which are slower, controlled, effortful, and make use of different information.)

If this hypothesis is correct, then it might seem that the best policy response would be to encourage people to engage in deliberative decision-making. However, it is not clear that such encouragement will succeed. Given our limited cognitive resources, it is sensible to make *some* decisions in a quick-fire way. Insofar as we do, we will be susceptible to projection bias.

In sum, although it seems unjustified to label projection bias "irrational", it seems right to ascribe it to lack of information or imperfect processing of information. The behaviour to which it gives rise is therefore not wholly voluntary in the aforementioned sense, so that there may exist a soft-paternalist justification for government action. Moreover, it may be that projection bias cannot be eliminated by providing individuals with information and encouraging them to engage in deliberative decision-making. Rabin may therefore be right that we have reason to counter its detrimental effects by taxing bad habits and subsidizing good ones.

3. Present bias and naïveté

A person who displays *present bias* gives extra weight to well-being *now* over any future moment, but applies the same discount factor to all future moments. This gives rise to time inconsistency in the agent's preferences. By way of illustration, consider the smoking habit described in Table 1 as it would be evaluated by someone who assesses the utility of an action as two times the utility it yields in the current period plus the sum of utilities in all future time periods. At t1, this person will evaluate smoking as preferable to not-smoking.[3] However, at all preceding and all subsequent times, he will evaluate smoking as inferior to not-smoking.

Rabin regards such time-inconsistent preferences as irrational. Once again, this description is not always apt. Present bias could be displayed by a person who, at every point in time, deliberates well on the basis of all relevant information and is fully in control of himself. He merely undergoes preference change which always leads him to give extra weight to the pleasures of the moment. Such a person always does what he *then* most prefers. In this case, even though present bias leads to lesser satisfaction of the individual's long-term preferences, the behaviour to which it gives rise cannot be regarded as irrational or non-voluntary (Parfit 1986, part 2). Taxing it therefore lacks soft-paternalistic justification. This does not mean a government has no liberal policy instruments at its disposal. For it can empower citizens to stop their future selves from acting on present bias. An example is the Australian requirement on gambling providers to give customers the option to self-exclude from their venue or products.

Of course, present bias may instead be due to familiar faults in reasoning or self-control. For example, a person might prefer to gamble because his deliberation is clouded by excessive desire, which leads him to give less credence than he should to evidence that he will lose a lot of money. Or he may arrive at the right conclusion through deliberation (that he ought not to gamble) but fail to abide by it because of weakness of will. In such cases, it is indeed appropriate to describe his present-biased choice as irrational and not fully voluntary. But even in such cases, a liberal government ought to prefer empowering citizens to control themselves (as the Australian policy does) to taxing them (as happens, for example, in Singapore, which heavily taxes gambling), at least where both are equally effective.

However, Rabin's analysis highlights a pitfall for policies that promote citizens' abilities to control their future selves. Such control depends on knowing that, in future, one will be biased towards the pleasures of the moment. Promoting such self-knowledge is therefore a component of an empowerment strategy. But, Rabin points out, such self-knowledge may be detrimental. Sophisticates, he writes, may be less likely than naïfs to start an advantageous good habit, because they regard the chances that they will stick with the good habit as lower than the naïfs. Self-knowledge may therefore be damaging.

To see why, imagine that present-biased Ahorita is considering embarking on an exercise regimen which would invariably do her good in the long run. Also suppose there is some uncertainty about the "start-up costs" of this habit: it is certain to be unpleasant at t1, but at t2 there is an equal chance that it is either no longer unpleasant (because her body adapts quickly), or *very* unpleasant (because her muscles are sore). Finally, suppose that if it is *very* unpleasant, her present bias at t2 would lead her to quit the regimen at t2, but if it is no longer unpleasant, she will stick with it. How will Ahorita decide at t1?

If she is *sophisticated,* she will predict that if she were to start exercising at t1, there is a good chance that she will drop out at t2 and have suffered at t1 for nothing. Moreover, at t1,

[3] The utility of smoking is 2×10+7+0=27; the utility of not smoking is 2×6+6+6=24.

she will weight the unpleasantness of exercise in that period disproportionately heavily. Together, her present bias and sophisticated scepticism will make it less likely that she will start the healthy habit.

By contrast, if she is *naïve* about her present bias, she will (falsely) predict that if she were to start exercising at t1, she would stick with it no matter what. Her naïveté will therefore lead her to *overestimate* the expected benefits of starting the exercise regimen. Of course, her present bias at t1 will also lead her to *overweight* its expected cost. Her naïveté therefore works against her present bias. Now, once she starts, she may drop out at t2. But there is also a good chance that she will stick with it, and this chance is sufficient (we can suppose) to make giving it a go worthwhile. Her naïveté therefore makes it more likely that she will start and stick with an advantageous pattern of behaviour.

What are the lessons for policy? First, from a soft paternalist point of view, the best policy is one which informs citizens about their present bias while ensuring they have access to effective commitment devices to control their present bias. Second, in the absence of such devices, helping citizens to see the truth about themselves may not be good policy, because it may neither make them better off nor promote autonomous pursuit of their long-term goals. Third, when present bias is irrational, effective commitment devices are not available, and education alone will not help, governments have reason to turn to Rabin's proposed taxes and subsidies.

Conclusion

Projection bias, present bias, and lack of self-knowledge make us less capable of fulfilling our long-term, considered desires. Respect for our rights of self-governance should lead us to favour policies that help us to overcome these biases by education and by facilitating self-binding, where these policies will be effective. But a careful look at the way these biases work indicates that these autonomy-enhancing strategies will not always be effective. In such cases, and only in such cases, governments have a soft-paternalistic justification for the use of taxes to counteract these biases.

References

De Marneffe, Peter. "Avoiding Paternalism," *Philosophy and Public Affairs* 68 (2006): 68-94.

Feinberg, Joel. *Harm to Self.* Oxford University Press, 1986.

Kahneman, Daniel. *Maps of Bounded Rationality*. Nobel Prize Lectures, 2002. On http://www.nobelprize.org/nobel_prizes/economics/laureates/2002/kahnemann-lecture.pdf

Parfit, Derek. *Reasons and Persons*. Oxford University Press, 1986.

Van Baal, Pieter, Johan Polder, Ardine de Wit, Rudolf Hoogenveen, Talitha Feenstra, Hendriek Boshuizen, Peter Engelfriet, and Werner Brouwer. "Lifetime Medical Costs of Obesity: Prevention No Cure for Increasing Health Expenditure," *PLoS Medicine*, 5.2 (February 2008): 242-9.