Studying Optimal Paternalism, Illustrated by a Model of Sin Taxes

By Ted O'Donoghue and Matthew Rabin*

The classical economic approach to policy analysis assumes that people always respond optimally to the costs and benefits of their available choices. A great deal of evidence suggests, however, that in some contexts people make errors that lead them not to behave in their own best interests. Economic policy prescriptions might change once we recognize that humans are humanly rational rather than superhumanly rational, and in particular it may be fruitful for economists to study the possible advantages of *paternalistic policies* that help people make better choices.

We propose an approach for studying optimal paternalism that follows naturally from standard assumptions and methods of economic theory: Write down assumptions about the distribution of rational and irrational types of agents, about the available policy instruments, and about the government's information about agents, and then investigate which policies achieve the most efficient outcomes. In other words, economists ought to treat the analysis of optimal paternalism as a mechanism-design problem when some agents might be boundedly rational.

This approach has many advantages. First and foremost, by *explicitly* addressing when and how people do and don't pursue their own best interests, economists will be better able to contribute to policy debates. To contribute to debates over regulating private financial decisions, we must study whether financial decisions are based on fallacious statistical reasoning and whether self-control problems lead people to borrow too heavily; to contribute to debates over teenage smoking, we must study whether teenagers become smokers against their long-run best interest. Economists will and should be ignored if we continue to insist that it is *axiomatic* that constantly trading stocks or accumulating consumer debt or becoming a heroin addict must be optimal for the people doing these things merely because they have chosen to do it.

A second advantage of our approach is that it forces us to look for the *best* feasible policy. Our

approach therefore imposes a check against promiscuous paternalism by public or private entities. Most adults in most situations make better choices for themselves than others would make for them, and a careful study of optimal paternalism will surely reinforce many of the traditional economic arguments against paternalism. Indeed, in some instances it will surely turn out that paternalistic policies do more harm than good (although even here our approach forces us to draw such conclusions through explicit analysis rather than *a priori* assumption).

Our approach also highlights how heavy-handed policy interventions are often inferior to less intrusive interventions, because heavy-handed interventions — such as banning purchases of goods that some people are prone to mistakenly consume — can cause significant harm to those for whom the behavior is rational. Concern about heavy-handed paternalism has already led some researchers to focus on minimally interventionist policies — e.g., O'Donoghue and Rabin (1999a) discuss "cautious paternalism"; Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin (2003) explore "asymmetric paternalism"; and in this session, Cass Sunstein and Richard H. Thaler (2003) investigate "libertarian paternalism". While these approaches differ, they all promote finding policies that help people who make errors while having little effect on those who are fully rational. Examples of such policies include simple education, small modifications of short-term incentives, or informed setting of easy-to-change default options.

But this brings us to yet another advantage of our approach: It reveals when this focus on minimal interventions might be misplaced. In some instances, even seemingly large deviations from the policy that is optimal for fully rational economic agents would not cause severe harm to those agents. In such cases, even a small probability of people making errors can have dramatic effects for optimal policy.

To illustrate our approach to studying optimal paternalism — and some of the principles above — we analyze a simple model of sin taxes for unhealthy goods.

I. Present-Biased Preferences and Optimal Sin Taxes

Economists currently analyze optimal commodity taxation based on an assumption that people's choices reflect their own best interests. As a result, economists have not investigated the case for sin taxes on unhealthy items based solely on the harm consumers may do to themselves.[1] Based on ongoing work, we show — using the same assumptions and tools as already used in public-finance theory — that imposing seemingly large sin taxes on unhealthy items while lowering taxes on other items may not hurt rational consumers by much (relative to the optimal "Ramsey" taxes). At the same time, such a policy change can create significant benefits for those who overconsume the unhealthy items due to self-control problems.

Most humans exhibit *present-biased preferences*: They have a tendency to pursue immediate gratification in a way that they themselves disapprove of in the long run. Recent research on such preferences — beginning with David Laibson (1997) — uses a simple and convenient functional form: A person's intertemporal preferences at time t are given by

$$U^t(u_t, ..., u_T) \equiv u_t + \beta \sum_{\tau=t+1}^{T} \delta^{\tau-t} u_\tau,$$

where u_τ is her instantaneous utility in period τ.[2] The parameter δ represents standard time-consistent impatience; for $\beta = 1$ these preferences reduce to standard exponential discounting. The parameter β represents a time-inconsistent preference for immediate gratification, where $\beta < 1$ implies an extra bias for now over the future. We and other researchers often refer to β as representing a "self-control problem", because it reflects a short-term desire or propensity that the person disapproves of at every other moment in her life. Our welfare analysis below treats this preference for immediate gratification as an error.

We analyze the implications of self-control problems in a simple consumption model of the form introduced by Frank P. Ramsey (1927). Consider a quasi-linear economy with three goods, potato chips, carrots, and a numeraire (perhaps best interpreted as leisure). Potato chips and carrots are

produced from the numeraire at a constant marginal cost of one. All markets are competitive, and we normalize the price of the numeraire to be one. The government raises revenue by taxing potato chips and carrots, and therefore the market price of these goods will be equal to one plus the commodity tax. (As in the optimal-taxation literature, we assume lump-sum taxation is not feasible; and as is well known, the government cannot do better by taxing sales of the numeraire.) Time is discrete, and production and consumption occur in all periods.

Consider first the case of a single, representative consumer. Let x_t, y_t, and z_t denote, respectively, her period-t consumption of potato chips, carrots, and the numeraire. We assume the person's instantaneous utility in period t is $u_t \equiv \rho \ln(x_t) + \sigma \ln(y_t) + z_t - \gamma \ln(x_{t-1})$, where $\rho, \sigma, \gamma > 0$ are exogenous parameters (to simplify our analysis, we shall assume throughout that $\rho \geq \gamma$ for all consumers). Carrots and the numeraire are "standard" goods in the sense that all utility effects of current consumption are experienced in the current period. Current consumption of potato chips, in contrast, creates negative health consequences in the future (the fact that it goes only one period forward is not essential). In period t, the person faces budget constraint $p_x x_t + p_y y_t + z_t \leq B$, where B is the person's endowment of the numeraire (available leisure time) in each period. We assume B is sufficiently large that $z_t > 0$ for all t.

Assuming $\delta = 1$ and $\beta \leq 1$, in period t the person will choose (x_t, y_t, z_t) to maximize $u^*(x_t, y_t, z_t) \equiv \rho \ln(x_t) + \sigma \ln(y_t) + z_t - \beta\gamma \ln(x_t) = (\rho - \beta\gamma) \ln(x_t) + \sigma \ln(y_t) + z_t$. The person will behave the same in all periods, and in particular will spend $\rho - \beta\gamma$ on potato chips and σ on carrots. Hence, in each period the demand functions are $x^* = (\rho - \beta\gamma)/p_x$, $y^* = \sigma/p_y$, and $z^* = B - (\rho - \beta\gamma + \sigma)$. By contrast, *optimal* behavior (that maximizes long-run well-being) maximizes $u^{**}(x_t, y_t, z_t) \equiv (\rho - \gamma) \ln(x_t) + \sigma \ln(y_t) + z_t$. Hence, in each period optimal behavior is $x^{**} = (\rho - \gamma)/p_x$, $y^{**} = \sigma/p_y$, and $z^{**} = B - (\rho - \gamma + \sigma)$. Observe that $x^* > x^{**}$ whenever $\beta < 1$; the self-control problem leads to over-consumption of potato chips.

Observed demand behavior is determined solely by σ and $\rho - \beta\gamma$, and not by the individual

components of $\rho - \beta\gamma$. Hence, a consumer with $(\sigma, \rho, \gamma, \beta) = (90, 110, 100, 1)$ or $(90, 105, 100, .95)$ or $(90, 100, 100, .9)$ would all have the same observed behavior — they would all consume \$10 worth of potato chips and \$90 worth of carrots each period. But only in the first case *should* the person consume \$10 worth of potato chips. In the second case, she should consume only \$5. In the third case, she shouldn't consume *any* potato chips. These examples illustrate a general lesson. If we were absolutely sure that a person has exactly zero self-control problem, we would know that her consumption of potato chips reflects the net benefits from consuming them. Otherwise, without further information beyond observed behavior, we know nothing about how sensible her consumption of potato chips is.

Now consider the government's choice of commodity taxes on potato chips and carrots when it must raise revenue R (per person per period). Given per-unit taxes t_x and t_y, the market prices for consumers will be $p_x = 1 + t_x$ and $p_y = 1 + t_y$. A representative consumer with parameter values $(\sigma, \rho, \gamma, \beta)$ will then consume $x^* = (\rho - \beta\gamma)/(1+t_x)$, $y^* = \sigma/(1+t_y)$, and $z^* = B - (\rho - \beta\gamma + \sigma)$. The optimal taxes are those that maximize $(\rho - \gamma)\ln(x^*) + \sigma\ln(y^*) + z^*$ subject to $t_x x^* + t_y y^* \geq R$. (Note that we account for β when predicting the response to taxes, but we — like the consumer herself in the long run — don't account for β in our welfare function.) It is straightforward to derive that the optimal taxes are:

$$\begin{aligned} t_x^* &= \frac{R}{\rho - \beta\gamma + \sigma - R} + \frac{\gamma(1-\beta)(\sigma/(\rho-\gamma))}{\rho - \beta\gamma + \sigma - R} \\ \text{and } t_y^* &= \frac{R}{\rho - \beta\gamma + \sigma - R} - \frac{\gamma(1-\beta)}{\rho - \beta\gamma + \sigma - R}. \end{aligned}$$

If there were no present-bias, so $\beta = 1$, we would have the standard conclusion that the two goods should be taxed equally (since they have identical elasticities). If for a fully rational consumer with $(\sigma, \rho, \gamma, \beta) = (90, 110, 100, 1)$ the government must raise \$5 in revenue from the \$100 total spent on potato chips and carrots, the optimal (Ramsey) taxes are $(t_x^*, t_y^*) = (\frac{1}{19}, \frac{1}{19})$ — about $5\frac{1}{4}\%$ tax on both items. But how would this consumer be affected by deviations from equal taxation? Suppose we raised the same \$5 in revenue by doubling taxes on potato chips and lowering taxes

on carrots. These taxes of $(t'_x, t'_y) = (\frac{2}{19}, \frac{17}{361})$ would, of course, hurt the consumer by distorting her consumption towards fewer potato chips and more carrots. But the harm from this distortion is small — less damaging than if 2 cents were taken away from the consumer. While this example is overly simple, we suspect that a similar conclusion would emerge from more realistic analyses: *According to rational-choice theory*, doubling taxes on potato chips is, at worst, not very damaging.

Consider instead a consumer with self-control problems who has $(\sigma, \rho, \gamma, \beta) = (90, 100, 100, .9)$. If again the government must raise \$5 in revenue from the \$100 total spent on potato chips and carrots, and if the government mistakenly assumed $\beta = 1$, it would choose the same $5\frac{1}{4}\%$ tax on both items. But such taxes would lead to overconsumption of potato chips. If instead we double taxes on potato chips as above, this consumer would be made better off, because higher potato-chip taxes help counteract the overconsumption of potato chips. In fact, the same change in taxes that would be like taking 2 cents away from the $(\sigma, \rho, \gamma, \beta) = (90, 110, 100, 1)$ consumer would be like giving an extra 47 cents to the $(\sigma, \rho, \gamma, \beta) = (90, 100, 100, .9)$ consumer. In other words, an increase in taxes on unhealthy items that has only a second-order cost to 100% rational consumers can have a first-order benefit to those with mild self-control problems. Hence, increasing taxes substantially on these items may be worth doing even if we suspect most people are very close to fully self-controlled.

To make this point more precise, we extend our model to heterogeneous consumers. Suppose consumers differ in terms of $(\sigma, \rho, \gamma, \beta)$, with average values $(\bar{\sigma}, \bar{\rho}, \bar{\gamma}, \bar{\beta})$. For simplicity, we further assume that the average value of $\beta\gamma$ is $\bar{\beta}\bar{\gamma}$ (which holds in our examples below). To investigate optimal taxation with heterogenous consumers, we must make interpersonal trade-offs. We put "equal weights" on each consumer. With these assumptions, one can show that, if the government must raise revenue R (per person per period), the formulas for the optimal taxes are exactly as above except for replacing the representative consumer's parameters $(\sigma, \rho, \gamma, \beta)$ with the average values $(\bar{\sigma}, \bar{\rho}, \bar{\gamma}, \bar{\beta})$.

Now consider a population with two types of consumers, fully rational consumers who have $(\sigma, \rho, \gamma, \beta) = (90, 110, 100, 1)$, and consumers with self-control problems who have $(\sigma, \rho, \gamma, \beta) = (90, 100, 100, .9)$. How does optimal taxation depend on the distribution of consumers? If, for instance, we believe 50% of consumers have self-control problems, then the optimal way to raise \$5 is with a tax of $(t_x^*, t_y^*) = (1, 0)$ — we should double the *price* of potato chips and eliminate the tax on carrots. Indeed, even if we believe only 10% of consumers have self-control problems, the optimal taxes would be $(t_x^*, t_y^*) = (\frac{3}{19}, \frac{4}{95})$ — we should still triple the taxes on potato chips. As suggested by our earlier analysis, increasing taxes substantially on sin goods may be worth doing even if we suspect most people are very close to fully self-controlled.

The example above involves a trade-off between helping $\beta < 1$ consumers and hurting $\beta = 1$ consumers. While we suspect this trade-off will be an integral part of any more realistic analyses, our focus on the (contrived) case in which observed behavior is identical for all consumers obscures the fact that revenue-neutral tax changes can sometimes help many $\beta = 1$ consumers as well. Indeed, it is easy to construct examples in which *all* consumers benefit from an increase in taxes on the unhealthy item (relative to equal taxation) — by improving the behavior of the $\beta < 1$ consumers while lowering the overall tax burden on the $\beta = 1$ consumers.

Although our analysis above imposes the realistic constraint of using linear per-unit taxes, more sophisticated schemes may in principle be feasible. The optimal tax schedule for a population of consumers with $(\sigma, \rho, \gamma, \beta) = (90, 110, 100, 1)$ is $(t_x^*, t_y^*) = (\frac{1}{19}, \frac{1}{19})$, and the optimal tax schedule for a population of consumers with $(\sigma, \rho, \gamma, \beta) = (90, 100, 100, .9)$ is $(\widehat{t_x^*}, \widehat{t_y^*})$ with $\widehat{t_x^*}$ is arbitrarily high and $\widehat{t_y^*}$ is arbitrarily close to $-\frac{1}{19}$ (given $\rho = \gamma$, we can raise arbitrarily close to \$10 in potato-chip taxes without harming the consumer, and hence can subsidize carrots by \$5). When both types coexist, any single tax schedule makes both types worse off relative to their respective optima. But suppose consumers were offered the *choice* between (t_x^*, t_y^*) and $(\widehat{t_x^*}, \widehat{t_y^*})$ (let's ignore implementation issues for the moment). What would they choose?

It is clear that the $(\sigma, \rho, \gamma, \beta) = (90, 110, 100, 1)$ consumers would choose taxes (t_x^*, t_y^*) for the same reasons as we derived them as optimal. Determining what the $(\sigma, \rho, \gamma, \beta) = (90, 100, 100, .9)$ consumers would choose requires more effort. If they are choosing taxes that will apply in the future, their choice will depend on their beliefs about future behavior, and therefore we must make an assumption about their prediction of their own future preferences. Two common assumptions in the literature are that people fully predict their future self-control problem — they are "sophisticates" — or that they (incorrectly) predict they will have no self-control problem in the future — they are "naifs".[3] While these two types often behave quite differently, in this context it happens that they will make the same choice: Both will choose the $(\widehat{t_x^*}, \widehat{t_y^*})$ schedule over the (t_x^*, t_y^*) schedule. Sophisticates have precisely the same preferences as "we" do over their future behavior (which serves as a reminder that "our" preferences should be equal to the long-run preferences of the agents themselves), and hence will choose this tax schedule precisely because it optimally overcomes their self-control problem. Naifs, while behaving the same in the end, will think very differently. They think that in the future they will be fully self-controlled — and not want to eat potato chips. Hence, they view the choice between $(\widehat{t_x^*}, \widehat{t_y^*})$ and (t_x^*, t_y^*) as a choice between a subsidy vs. a tax on the only food, carrots, they will want to consume. Obviously, they will choose the subsidy.

This extreme result of perfect sorting into the type-specific optimal tax schemes is clearly not robust to more heterogeneity in preferences. But in general there may be room for improvement over the imposition of a single tax schedule. A bigger issue is whether such schemes can be implemented in the marketplace. Perhaps we could require consumers to choose one of two electronic cards needed for all purchases that would determine their tax schedules (which is becoming more feasible as more and more consumer transactions are carried out with cards of various stripes). Alternatively, perhaps we could require consumers to buy non-refundable coupons in advance that give them the right to purchase these items. If these coupons could be bought in any quantity, they would merely require a bit of foresight for 100% self-controlled consumers. The rest of us,

whether sophisticated or naive, may benefit — e.g., those whose New Year's resolutions are to eat fewer calories or quit smoking would not buy the coupons for potato chips and cigarettes.

We can even speculate about more efficient mechanisms. Suppose it were feasible to charge well-calibrated *sin licenses* whereby people pay a one-time (or few-time) fee for the right to purchase an item. In the example above, for instance, we might raise the same amount of revenue by charging a little less than \$5 per day for a license for either potato chips or carrots, and a little more than \$5 per day for a license for both goods. The $(\sigma, \rho, \gamma, \beta) = (90, 110, 100, 1)$ consumers would purchase the license for both items, and end up consuming $(x, y, z) \approx (10, 90, B - 105)$. The $(\sigma, \rho, \gamma, \beta) = (90, 100, 100, .9)$ consumers would purchase solely the carrot license, and end up consuming $(x, y, z) \approx (0, 90, B - 95)$. Both types are better off than they would be under any per-unit taxation scheme — indeed, in this example, such licenses achieve the first best. (For fully rational consumers, we achieve the first best merely because the licensing scheme is equivalent to lump-sum taxation; for consumers with self-control problems, in addition the licensing scheme eliminates undesirable consumption of potato chips.)

Applying these principles outside of our framework, suppose that instead of charging a \$2-per-pack tax on cigarettes, we charged \$5000 for a picture I.D. that allows that person to purchase up to 2500 packs tax free — and made it illegal to purchase cigarettes without this I.D. Such a policy change could help *all* types of consumers. All the 18-year-olds who are rationally deciding to become lifetime nicotine addicts would purchase the license. The 18-year-olds who instead end up paying \$5000 in taxes for a lifetime habit they did not identify as optimal when they started would not buy the licenses. (If there were concerns that this scheme would prevent optimal experimentation, we could also issue a one-time "learner's permit" allowing a person to purchase up to 10 packs of cigarettes.)

II. Discussion

Our analysis is not at all conclusive on the particulars of the proposed taxes; it is only suggestive of the types of questions we can start to ask once we abandon the *a priori* assumption of 100% rationality. More generally, we are hesitant to advocate vast policy change based on models from behavioral economics that are still at an early stage of development.

There are many reasons that paternalistic policies might be undesirable — such as fears of regulatory capture or transactions costs in implementation. While we do not address such issues, we are sympathetic to them. Even so, we feel that they ought to be articulated carefully and precisely, and investigated, rather than loosely invoked as pretexts for the blanket rejection of paternalism.

A more general concern with our approach is that the efficacy of proposed paternalistic policies can depend critically on having identified all the plausible mistakes people can make. A policy that helps those agents making one common error may *hurt* those making another common error, even if the policy is virtually harmless to fully rational people. Researchers studying optimal paternalism must address such possibilities.

Despite these reservations, we are even more hesitant to continue to make policy prescriptions based solely on the axiom of 100% rationality. The possibilities that 15-year-olds err in becoming tobacco addicts or that 25-year-olds err in borrowing heavily on their credit cards or that 35-year-olds err in too wildly playing the stock market with their retirement savings all strike us as profoundly plausible and of great policy relevance. It therefore seems to us that policy analysis that incorporates the substantive insights and methodological rigors of economics, while being more realistic about the nature of errors people make, should be enthusiastically and quickly embraced.

References

Camerer, Colin; Issacharoff, Samuel; Loewenstein, George; O'Donoghue, Ted; and Rabin, Matthew. "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism'." Penn Law Review, forthcoming.

Frederick, Shane; Loewenstein, George; and O'Donoghue, Ted. "Time Discounting and Time Preference: A Critical Review." Journal of Economic Literature, June 2002, 40(2), pp. 351-401.

Gruber, Jonathan and Koszegi, Botond. "A Theory of Government Regulations of Addictive Bads: Optimal Tax Levels and Tax Incidence for Cigarette Excise Taxation." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 8777, February 2002.

Gruber, Jonathan and Mullainathan Sendhil. "Do Cigarette Taxes Make Smokers Happier?" Mimeo, Massachusetts Institute of Technology, 2002.

Laibson, David. "Golden Eggs and Hyperbolic Discounting." Quarterly Journal of Economics, May 1997, 112(2), pp. 443-477.

O'Donoghue, Ted and Rabin, Matthew. "Procrastination in Preparing for Retirement," in Henry Aaron, ed., Behavioral Dimensions of Retirement Economics. Washington DC and New York: Brookings Institution Press and Russell Sage Foundation, 1999a, pp. 125-156.

_____. "Doing It Now or Later." American Economic Review, March 1999b, 89(1), pp. 103-124.

_____. "Choice and Procrastination." Quarterly Journal of Economics, February 2001, 116(1), pp. 121-160.

Ramsey, Frank P. "A Contribution to the Theory of Taxation." Economic Journal, March 1927, 37(145), pp. 47-61.

Sunstein, Cass and Thaler, Richard H. "Libertarian Paternalism." American Economic Review, May 2003 (Papers and Proceedings).

Footnotes

* Department of Economics, Cornell University, 414 Uris Hall, Ithaca, NY 14853-7601, and Department of Economics, University of California, Berkeley, 549 Evans Hall #3880, Berkeley, CA 94720-3880. Some of the ideas here are based on joint work with Colin Camerer, Sam Issacharoff, and George Loewenstein. For financial support, we thank the National Science Foundation (grants SES-0214043 and SES-0214147), and Rabin thanks the Russell Sage and MacArthur Foundations.

1. Two recent exceptions are Jonathan Gruber and Botond Koszegi (2002), who study optimal taxation of cigarettes, and Gruber and Sendhil Mullainathan (2002), who provide some empirical evidence that cigarette taxation increases welfare.

2. For a recent overview of the many contributions to this literature, see Shane Frederick, Loewenstein, and O'Donoghue (2002).

3. See O'Donoghue and Rabin (1999b, 2001) for a discussion of sophistication and naivete, and an approach to modeling the intermediate cases of "partial naivete".