CHAPTER 30

Paternalism, Manipulation, Freedom, and the Good

JUDITH LICHTENBERG

The creature who has come to be known as homo economicus differs from living, breathing human beings in two central ways. First, homo economicus is fully rational: he always employs means that maximize the fulfillment of his ends and does what is in his best interests.[1] Human beings are often not rational; as a result of cognitive errors and biases, emotional reactions, and volitional weaknesses, they often fail to act in their own best interests. Behavioral economists and psychologists in this book and elsewhere have greatly increased our understanding of how human beings fall short in these respects and what can be done to more closely align their means with their ends (e.g. Barr, Mullainathan, and Shafir; Mullainathan and Shafir; Ubel; Wansink; all in this volume).

Human beings differ from homo economicus also in the ends they seek. Economists and others have often construed people's ends in terms of their narrow self-interest, particularly their economic self-interest. Yet—as careful thinkers note, although sometimes only when pushed—nothing in economic theory dictates the content of a person's ends or preferences. Assume an agent as altruistic as you please, whose deepest desire is to eliminate suffering and disease in the world. The fallacy of thinking agents must be self-interested results from confusing the *subject* of my preferences (me) and the *object* of my preferences (often me, but sometimes others) (Lichtenberg 2008, 2010a).[2] Several authors in this book acknowledge such other-regarding or altruistic preferences, which they call "social motivations" (Tyler, this volume; Weber, this volume). They suggest that we should capitalize on such motivations or try to enlarge their influence on behavior.

So behavioral economists and psychologists have called into question both assumptions about homo economicus—that he is rational (employs the optimal means to his ends) and that he is self-interested (cares only for his own well-being). But the two challenges pull in opposite directions. If human beings are less rational than homo economicus, then clearly they fall short. Their human traits constitute defects we should try to remedy or counteract, if we can do so without introducing other problems that are worse.[3] But if human beings are not (necessarily) self-interested, that is a good thing. Or so I shall assume. Homo economicus, then, is in one way worse and in another way better than real human beings.

Rationality and the Good

What does it mean to say that people sometimes act less than rationally? One way to understand the claim is to say that they fail to do what is in their best interests or to realize their own Good. For example, they fail to save adequately for retirement; they eat too much or unhealthily; they do not take their medicines as they should. But talking about a person's best interests immediately raises the question: best interests according to whom? In liberal societies it is natural to parse this concept in terms of an agent's own desires or preferences. In a common formulation, a person's best interests are what she would want if she possessed full information and suffered no cognitive, emotional, or volitional defects and biases. Such a definition might not always produce a determinate answer to the question of what is in a person's best interests, but we can suppose it does at least some of the time.

Of course, what people *would* want under these ideal and unrealizable circumstances is not equivalent to what they *do* want. That is part of the problem. But it may be less misleading to acknowledge that people have various wants and preferences that sometimes conflict. They want a comfortable retirement but also prefer more income now; they prefer to be fit and healthy but also like ice cream. Often such conflicts can be understood in terms of the distinction between short-term and long-term preferences. We can also distinguish levels or orders of preferences:

a smoker may have a first-order desire to smoke and a second-order desire not to smoke—that is, a desire not to desire to smoke. In any case, to say that people sometimes act less than rationally is to suggest that their desires can be ranked, preferably from their own point of view as well as from others'.

It is rarely helpful, however, to talk about what a person *really* wants, which suggests that although the person behaves as if she wanted one thing, in truth she wants something else. People's desires are multiple and conflicting. There is no plausible way to escape the conclusion that people hold inconsistent wants, desires, or preferences, and we should avoid linguistic tricks that seem to make these inconsistencies disappear.

Paternalism, Hard and Soft

Forcing people by law or some other form of regulation to act in their own best interests has traditionally been called *paternalism*. But several years ago Sunstein and Thaler introduced the concept of libertarian paternalism, which "attempts to influence the choices of affected parties in a way that will make choosers better off" without forcing them to do something or refrain from doing something (Sunstein and Thaler, 2003, p. 1162). So now we distinguish between classical and libertarian paternalism, or hard and soft paternalism.

It might seem almost a truism to say that if you can get people to change their behavior for their own good without forcing them, that is better than bringing the long arm of the law down on them. Yet although most people in liberal societies would prefer to avoid paternalism, there are probably irreducible differences in people's tolerance for it. Political libertarians think the price is always too high. Perhaps it is less misleading to say that they oppose it on principle. Others disagree; they think that the benefits of paternalism sometimes outweigh the costs.

Still, most people probably agree that we should minimize the use of coercion in guiding people to do what is good for them. We will be least uneasy if they choose freely and knowledgeably what is best for them. Alas, it turns out that information is not enough (Ubel, this volume). So the question is whether and to what extent we can induce (entice? cause?) people to do what is best for themselves—or, for that matter, others—without forcing them.

The validity of soft paternalism rests on at least two assumptions. One is that we can somehow formulate a coherent idea of a person's best interests, their Good—for example, in terms of what satisfies their long-term or higher-order preferences—and that it is better, other things being equal, if people achieve their Good than if they do not. It is not necessary that we be able to give a complete account of what is in a person's best interests, as long as we can give a determinate account in some cases.

The other assumption is that, as Thaler and Sunstein put it, there is no such thing as neutral design: every environment exhibits features—a "choice architecture"—that nudge agents in some direction rather than others, making it more likely that they will do X rather than Y or Z. A different way of putting the point is that human behavior is "heavily context dependent" (Barr, Mullainathan, and Shafir, this volume). In the psychological literature the technical term for this view is *situationism*, which insists on the power of situational factors over individuals' personal traits to determine behavior. Cafeteria items may be arranged in a variety of ways, but they must be arranged somehow, and their order may significantly influence people's food choices and thus their health (Thaler and Sunstein, 2008; Thaler, Sunstein, and Balz, this volume). Those who serve food must place it on plates of some size or other; plate size affects how much people eat (Wansink, this volume). Doctors must explain treatment options to their patients in some order, using particular language, and expressing probabilities in a particular way (McNeil et al., 1982; Ubel, this volume). Employers, governments, and others who offer policies regarding retirement benefits, insurance, organ donation, and other matters can offer opt-in or opt-out defaults (Johnson and Goldstein, 2003, this volume). These decisions may have profound effects on people's choices and thus on their well-being.

Paternalism and Manipulation

I want to make several points about the distinction between hard and soft (or traditional and libertarian) paternalism. First, as Thaler and Sunstein acknowledge (2008), the distinction is not sharp, since one can choose to violate even legally coercive rules, accepting the penalty or (more likely) taking the risk that one will not be caught. It does not follow that the distinction between legally coercive rules and other forms of influence is trivial, but we should note that influence is a matter of degree, with many points along the continuum between liberty and force.

Still, it is natural to think that not forcing people to act (or not act) is preferable to forcing them; better to leave the choice more open even if influence is inevitable. Yet in one way coercion might be preferable: it is overt and explicit. Citizens know that the state is attempting to control them when it prohibits riding

a motorcycle without a helmet. But they are likely not to notice the significance of the arrangement of food in the cafeteria or its influence on our behavior. Similarly with the default choice of retirement plans and other policies. The idea that someone is attempting to influence our choices without our knowledge or consent is troubling and may seem in some way at least as much a violation of our liberty as explicit coercion. We tend to call this kind of influence-creation *manipulation*; its connotations are negative.

One might respond that this objection neglects the idea that some arrangement or other of the choice environment is *inevitable* and that there is no neutral design. In this section I consider one aspect of this response; in the next section, another.

Suppose that nonneutrality is indeed inevitable. Still, manipulation might be reduced if policy makers were required to reveal more clearly how they attempt to influence decisions, so that agents could more easily resist their influence if they so chose. Of course, we know from behavioral economists and psychologists that awareness and knowledge are not always enough. Sometimes the difficulty is rather in the link between intentions (formed in light of knowledge)—with which the road to hell is paved—and action, as Barr, Mullainathan, and Shafir (this volume) argue.[4]

At the very least, designers of defaults can sometimes control how easy or hard it is to depart from them. For example, mortgage rules can be structured with opt-out defaults that "make it easier for borrowers to choose a standard product" and harder to choose one they are less likely to understand or to be able to afford (Barr, Mullainathan, and Shafir, this volume). Yet in many contexts transparency is unrealistic or impossible. Must the cafeteria managers explain the reason for their food arrangement or for the size of their plates? Must the Motor Vehicle Administration explain why it uses an opt-out rather than an opt-in default for organ donation? Transparency may be useful in some contexts but not in others.

Defaults

The second response to the claim that there is no neutral design is to question it outright. Consider the example of defaults, which seem to illustrate the nonneutrality thesis. Johnson and Goldstein (2003, this volume) have shown the profound effects of defaults on organ donation and other policies. Although organ donation is not a matter of paternalism but of other-regarding choices (about which I say more below), the mechanisms are the same as for paternalistic intervention.

In some countries, including the United States and Great Britain, you must choose (when you get or renew your driver's license) to become an organ donor; the default is not to donate. In many European countries, the policy is the reverse: consent to donating one's organs is presumed, and one must explicitly opt out to avoid donation. In Austria, France, Hungary, Poland, and Portugal, which all have opt-out policies, effective consent rates are over 99%. In countries with opt-in policies, consent rates are radically lower—from 4.25% in Denmark to 27.5% in the Netherlands.[5]

Yet a no-default policy is also possible: forced or mandated choice. In an online experiment, Johnson and Goldstein (2003) show that mandated choice approximates the opt-out default: 79% of participants who must decide choose to be organ donors; 82% in the opt-out default remain as donors; only 42% in the opt-in condition agree to be donors.

Are mandated choices counterexamples to the claim that neutral design is impossible? To fully answer this question would require an extended inquiry into the nature of neutrality, and even after it, we might still not reach a clear or uncontroversial answer. What seems certain is that mandated choice is *more* neutral than opt-in or opt-out defaults.

But that is not the end of the matter, because neutrality is not the only value and may not always be the most important one. Thaler, Sunstein, and Balz (this volume) argue that where choices are difficult or complicated, people may prefer a "good" or "sensible" default; and when choices are not binary, yes-no decisions, mandated choice might not even be feasible.

What is a good default? Perhaps it is the one I would prefer if I had full information and sufficient time and mental resources to process it. Since people have different values and preferences, on this criterion no default is necessarily best for everyone. Some people would like to donate their organs, but some object on religious grounds. So the good default might be the one that most people would prefer. In the case of organ donation, Johnson and Goldstein's online experiment suggests that opt-out policies are better because they more closely match people's preferences when no default is offered. Even apart from cases where choices are not binary, however, it is implausible to think that people have preexisting preferences in many situations in which defaults are common and desirable (Thaler and Sunstein, 2003, pp. 1173–1174). I may not have a preference concerning the details of my software installation, even armed with full information and adequate mental resources. More serious still is that our preferences are partly constructed out of the choice situations in which we find ourselves and thus cannot be employed to structure those choice situations.

What can we conclude from this discussion? First, even if we agree that there is no neutral design, some

designs may be more neutral than others. But, second, neutrality does not always trump all other values. Especially if the aim is to do what is in people's best interests or satisfy their (deeper? more important? more enduring?) preferences, we will sometimes want to structure environments in ways that are in tension with choices they might otherwise make.

Politics, Power, and Freedom

One of Thaler and Sunstein's central aims seems to be to reassure those who worry about bringing the state's power down on individuals through paternalistic legislation that such crude techniques are not necessary. But the message of their work, and that of other behavioral economists and psychologists, might be seen in less rosy terms, a glass half empty rather than half full. Despite the desire to preserve freedom that leads us to resist hard paternalism, we are not very free at all. Subject to error, bias, ignorance, temptation, passion, and weakness of will, we find ourselves (or, more often, fail to realize that we are) buffeted about by the winds of influence, internal and external, intentional and accidental, self-interested and benevolent. We can learn to control some of the forces acting upon us so that we are better able to realize our Good, but sometimes it may seem not much more than a rhetorical trick to say we are thereby free.

Despite the wealth of insights behavioral economists and psychologists have provided, with a few exceptions there is a peculiarly apolitical quality to their work. One might infer from the literature that the cognitive, affective, and volitional deficiencies that lead agents astray are merely unfortunate natural facts; one might fail to see how they are actively exploited and encouraged by banks, insurance and credit card companies, fast-food conglomerates, and others who profit from these weaknesses. Altering the choice architecture so as to nudge people to serve their own best interests is important. But some entities need more than nudges. The activities of corporations and others who prey on individuals need nonpaternalistic, other-regarding restrictions, in addition to positive requirements that they serve individuals' interests (for important examples see e.g., Barr, Mullainathan, and Shafir, this volume; Mullainathan and Shafir, this volume). This is less a matter of behavioral economics in the usual sense than of the realities of politics and power.

Rationality and Morality

As I noted at the outset, economists and other social scientists often shrink from assuming that people are capable of acting altruistically. That reluctance may derive from a belief in egoism as a kind of default—the uncontroversial view that needs no defense and that keeps social science away from the dangerous territory of "value judgments." Yet the clear implication of behavioral economics and psychology (not to mention philosophy) is that we cannot avoid making value judgments. If there is no neutral design of choice environments, or if even the choice of a neutral design is itself not neutral (as the discussion above of reasons against no-default choices suggests), we have no alternative but to shape choice environments in accordance with some values or other. We should do so in accordance with a conception of what is genuinely in people's best interests or which preferences it is most important for them to satisfy. To leave the environment as it is (whatever that might mean) is also to make a value judgment, and the jumble of people's conflicting desires and preferences forces us to favor some and not others.

From the recognition that we need a conception of a person's good it is not much of a step to the conclusion that we need a conception of the general good. The value judgments inherent in the general conception are no more significant than in the individual, the gap between my immediate preferences and my best interests no wider than the gap between my good and your good.[6] Two other facts lend support to the legitimacy of taking into account more than people's egoistic choices. One is that, as others in this volume have argued (Tyler; Weber), individuals have social motivations: they care not only about themselves but also about others. In other words, they are somewhat altruistic (some more than others, of course).

The other is that nudging individuals to act in accordance with the interests of other people is rooted not only in the assumption that they would so choose but also in the fact that they have moral responsibilities to do so. The more minimal defense of these responsibilities rests on so-called negative duties. When our actions contribute to harming other people (or creating "externalities," as economists like to put it), those harmed may have valid claims against us; in many such cases the state is entitled or perhaps even required to enforce such claims. This much even political libertarians admit! Attempts to induce people to behave in ways less harmful to the environment can be rooted in these negative duties. Somewhat more controversial is the idea that we have not only negative duties not to harm others but also, at least sometimes, positive, "humanitarian" duties to help them. But how controversial is this view really? Do we need fancy arguments to be convinced that it would be better if people did not ignore genocide and other atrocities and that it is therefore legitimate to shape environments in ways that cause them to act accordingly (Slovic et al., this volume)?[7]

Notes

1. Whether these are equivalent is an open question I address briefly in what follows.

2. For experimental evidence of unselfish motives see, e.g., Batson (1991); Fehr and Fischbacher (2004). For most purposes the existence of unselfish motives is pretty obvious, but at a deep level, the claim is difficult to test, as Batson acknowledges. He and his colleagues attempt to test it through a number of complex experiments, all of which confirm the existence of altruistic motivations. As Sober and Wilson note (1998), however, this does not prove that all versions of egoism will fail. Because sophisticated versions appeal to the internal rewards of helping others—rather than simply money, say—it is always possible that a more subtle psychological reward lurks that the experiments have not detected (pp. 271–273). This possibility will strike many as far-fetched, confirming their suspicions that egoism is unfalsifiable, but it permits those attracted to egoism to hang on to their convictions.

3. Perhaps not all such differences between homo economicus and real human beings should be construed as defects. I leave that question aside, assuming only that at least some of these traits are flaws.

4. They discuss changes to the Truth in Lending Act that require credit card companies to disclose to customers information about the expected time it will take to pay off credit card balances if they pay only the minimum balance, and they argue that "such disclosures may not be strong enough to matter. . . . In fact, the borrower would need to change behavior in the face of strong inertia and marketing by credit card companies propelling her to make no more than the minimum payments."

5. Johnson and Goldstein offer three (non–mutually exclusive) explanations for the power of defaults: effort, implied endorsement, and loss aversion. Sunstein and Thaler (2003) suggest another important one: the idea that the default is "what most people do, or what informed people do" (p. 1180). This might appear similar to implied endorsement. But there are two possible differences. First, Johnson and Goldstein's idea focuses on the policy maker's endorsement, Sunstein and Thaler's on the public's. Second, an agent may choose what she believes is the popular choice not because people's choosing it signifies approval of some independently valuable good but simply because she wants to do what others are doing, irrespective of whether it has independent merit.

6. For a view showing the similarities between prudential and moral reasons see Nagel (1970).

7. For an argument that the distinction between negative and positive duties—between the duty not to harm and the duty to render aid—is exaggerated, see Lichtenberg (2010b).

References

Batson, D. (1991). *The altruism question: Toward a social-psychological answer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Fehr, E., and Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185–190.

Johnson, E. J., and Goldstein, D. G. (2003). Do defaults save lives? *Science, 302*(5649), 1338–1339.

Lichtenberg, J. (2008). About altruism. *Philosophy and Public Policy Quarterly, 28*(1–2), 2–6.

———. (2010a, October 19). Is pure altruism possible? *New York Times.* Retrieved from http://opinionator.blogs.nytimes.com/2010/10/19/is-pure-altruism-possible/

———. (2010b). Negative duties, positive duties, and the "new harms." *Ethics, 120*(3), 557–578.

McNeil, B. J., Pauker, S. G., Sox, Jr., H. C., and Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine, 306*(1), 1259–1262.

Nagel, T. (1970). *The possibility of altruism.* New York: Oxford University Press.

Sober, E., and Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior.* Cambridge, MA: Harvard University Press.

Sunstein, C. R., and Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review, 70*(4), 1159–1202.

Thaler, R., and Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth and happiness.* New Haven, CT: Yale University Press.

The Behavioral Foundations of Public Policy

EDITED BY ELDAR SHAFIR

Princeton University Press
Princeton and Oxford

2013