SEVEN

Utilitarianism, Consequentialism, and Justice

Consequentialism is the doctrine that one should judge things morally by their intrinsic value and by the value of their consequences. It specifies a particular structure for ethics. In a consequentialist framework one must first decide what is intrinsically valuable. Questions of intrinsic value are not necessarily the most important moral questions, but they must be answered first because everything else depends on their answers. Then one assesses actions, policies, and institutions in terms of their "results" – that is, their own value and the value of their consequences. Welfare economics presupposes a consequentialist moral theory in which only welfare has intrinsic value.

Consequentialist assessment is always comparative: an action, policy, or institution is morally right or permissible if its net results are no worse than the results of any alternative. If the results of a particular policy are better than those of any alternative, then the policy is morally *obligatory*. Whether a policy or action is right or wrong depends on both what state of affairs will obtain if it is implemented and what the state of affairs will be if any feasible alternative is implemented. Nothing is absolutely impermissible. The right action may have terrible consequences when the consequences of all of the alternatives are even worse, and even a policy with terrific results is impermissible if there is a better alternative.

A utilitarian is a consequentialist who says that what is intrinsically good is individual "welfare" or "well-being." Is individual welfare a mental state like happiness, the satisfaction of actual preferences, the satisfaction of "rational" or "informed" preferences, or something else not tied to preferences or mental states? Different utilitarian theories disagree about what well-being is.

The fundamental principle of utilitarianism is that one should do whatever maximizes overall welfare, taking everyone into account. This

formulation masks a disagreement between those utilitarians who favor the maximization of total welfare and those who instead support maximizing average welfare. However, both average and total utilitarians agree that the distribution of welfare does not matter: all that matters is its total or average amount.

Adding up or averaging well-being requires measuring the well-being of different individuals on the same scale. One must be able to compare the increases or decreases in one person's well-being to the increases or decreases in the well-being of others. Notice that utilitarianism is not a selfish doctrine. Morality demands that one maximize total or average welfare, not one's own welfare. Indeed, an important objection to utilitarianism is that it may require that people make impossibly large sacrifices to benefit others (see Murphy 2000).

How does utilitarianism work? Consider how a utilitarian would approach the question of whether there should be a significant estate tax. Since the entire focus of utilitarianism is on the consequences of policies for well-being, questions such as whether people have a fundamental right to dispose of their wealth as they wish are irrelevant except insofar as they contribute to well-being. So too are questions of moral desert or the weight of past promises. The only relevant question is whether imposing an estate tax (at some particular tax rate) results in more total or average welfare than the alternatives. So a utilitarian argument for or against a 50 percent tax on estates greater than $2,000,000 will depend on factual issues such as whether the tax enhances or constrains economic growth and the solvency of government. Everyone's welfare counts in this assessment, including both those upon whom an estate tax has a direct effect and the much larger number who are affected only indirectly. In the terms of what Mill calls "Bentham's dictum," "Everybody to count for one, nobody for more than one" (Mill 1863, ch. 5).

7.1 Clarifying Utilitarianism

This sketch of utilitarianism leaves many questions unanswered. First, and most crucially, what exactly is welfare? Individual "welfare" has many different meanings. Consider: are happiness and the satisfaction of preferences the same thing? Satisfying an agent's preferences does not always make the agent happier. (Think of those who seek happiness in a bottle of whiskey.) The major nineteenth-century utilitarians (especially Bentham, Mill, and Sidgwick) took utility to be a mental state like happiness or pleasure (or, more precisely, to be that property of objects that causes such mental states;

see Broome 1991a), but few contemporary moral philosophers agree. Most contemporary utilitarians take welfare to be the satisfaction of rational and informed preferences. Chapter 8 will explore views about what constitutes individual welfare.

Second, the consequences of actions, policies, and institutions are usually uncertain. When legislators instituted a policy believing mistakenly that its consequences would be better than any of the alternatives, did they do the right thing because the expected consequences were better than the alternatives, or did they do the wrong thing because the consequences of some alternative would have been better? It is awkward to take the first alternative and judge that the policy was "right from their ex ante perspective" but "wrong from an ex post perspective." It makes more sense to judge that according to the utilitarian standard the legislators made the wrong choice – no matter how rational, blameless, or even noble they were in making the mistake. In other words, what matter to the utilitarian assessment of an action or policy are its consequences, full stop. When we say to someone who blamelessly made a bad choice, "You did the right thing," we mean only that, given what the agent knew, she chose as well as she could – not that there was no alternative that would have been morally better. Although utilitarian appraisals of proposed actions and policies must be made ex ante and thus always depend on beliefs about what their consequences will be, whether the actions or policies are right depends on what the consequences turn out in fact to be.

Third, *whose* welfare counts? Should an agent consider only the welfare of currently living human beings? What about the welfare of those not yet born – and of those who might, as the result of one policy or another, not even be conceived? (See Parfit 1984, ch. 17.) These issues are particularly pressing with respect to the problem of climate change, because the policies adopted now are likely to have major effects on future generations. Climate change raises questions not only about how to take account of the well-being of individuals whose very existence may depend on our policies, but also concerning the fairness of our treatment of them. (These issues concerning the interests of future generations are as worrying for standard welfare economics as they are for utilitarianism.)

In addition to deciding which human interests count, there is also the question of whether evaluations of alternative policies should consider the welfare – or at least the pains and pleasures – of nonhuman animals. Most people believe that it is morally wrong to inflict suffering on animals "needlessly." Commercial farming of animals for their meat and hides raises moral questions, particularly when animals are confined in tight cages,

frightened, and badly treated. If utilitarians count the pleasures and pains of all sentient beings, then they can explain quite naturally why the treatment of animals matters morally (Singer 1975). But utilitarians face difficult problems comparing human and nonhuman welfare (how many frogs is a human life worth?) as well as measuring animal welfare. (What is the value to a chicken of being able to roam freely?) Beyond nonhuman animals, there are questions about how to capture the value of trees, rivers, or natural beauty: are these only valuable because of the pleasure they give to sentient beings capable of pain and pleasure?

Fourth, should utilitarians be concerned with total welfare or average welfare? Since average welfare is total welfare divided by the size of the population, total and average utilitarianism will coincide when population size is fixed. But if alternative policies have consequences for how many people there will be, then what maximizes total welfare may differ from what maximizes average welfare. Average utilitarianism seems initially more appealing. A small population whose average well-being is high seems more attractive than a much larger population with low levels of individual well-being, which has more total well-being. At the same time, how can it be wrong (as average utilitarianism implies) to do something that adds to total well-being, such as having a child whose life will be happy, although less happy than the average? These theoretical issues bear directly on pressing problems concerning population control (Parfit 1984, pt. 4; Broome 2004).

Fifth, how should utilitarianism guide individual and social decision making? Actions and policies are morally obligatory if they maximize utility, but it does not follow that the best way to make moral decisions is to attempt to calculate the welfare consequences of different actions and policies. Not only is calculating the consequences of actions on the welfare of the whole present (and future?) human (and animal?) population of the world fraught with uncertainties (and not much fun), it is also likely to introduce lots of bias because people tend to take a more favorable view of the overall consequences of actions that benefit them personally. Calculation is a time-consuming action, and not one that is likely to maximize utility. A person does not usually become happy by directly aiming to maximize his happiness; nor is he likely to do right by directly aiming to maximize total happiness. Furthermore, uncoordinated actions by numerous individuals may lead to bad aggregate consequences that could be prevented by requiring people instead to follow simple rules. People are more likely to perform actions that actually maximize welfare if they do not calculate the welfare consequences of their actions and instead act on general rules, such as "tell the truth," "keep your promises," and so forth.

Policy makers may be in a better position to investigate the consequences of alternatives, but they are also often better advised to adhere to rules, for it is difficult to tell what the consequences of policies will be and what impact they will have on total welfare. Consider, for example, the Bush and Obama administration policies of imprisoning "enemy combatants" indefinitely without any right to an attorney. Since the consequences of a policy such as this one are highly uncertain and could be disastrous, there is a utilitarian case for treating it as off-limits – which is precisely the point of constitutional and international law.

At this point it might appear that utilitarianism is pulling a disappearing trick, having taken its bow. It seems that the utilitarian is now ceding the stage to defenders of traditional moral principles. But even though the best way for individuals to maximize utility is usually to stick to traditional moral rules, utilitarianism does not endorse all of traditional morality. It seems plausible that rules against lying maximize utility, but it is not obvious, for example, that utilitarianism supports some of traditional morality's objections to expanding the scope of markets to permit paid adoptions ("baby selling") or buying and selling of organs for transplantation. People's sentiments do not always guide them well, and it is crucial that some people think carefully about what sorts of policies will do the most good and advise the rest of us. For example, defenders of "effective altruism" have argued that there are good ways to amplify the impact of charity (MacAskill 2015; Singer 2015).

In pointing out that people generally do better following established moral practices rather than attempting to estimate the aggregate welfare consequences of alternative actions, it may appear that we are endorsing "rule utilitarianism" (Hooker 2001; Mason 1998), but this is not the case. According to rule utilitarianism, actions are right if and only if they conform to a set of rules such that the consequences of the general adoption of these rules for total or average welfare are no worse than the consequences of the general adoption of any other set of rules. Hence rule utilitarianism maintains that people ought to adhere to the utility-maximizing rules even in the unusual case in which they are confident that breaking a rule would result in more total welfare. If the point of morality is to maximize welfare, then in these unusual circumstances, individuals should violate the rules.[1]

The sixth question concerns interpersonal comparisons of well-being and deserves a section to itself.

[1] Rule utilitarianism may, however, have other justifications. See Hooker (2001, ch. 4) and Parfit (2011).

7.2 Interpersonal Comparisons of Well-Being

If policy *A* benefits Ira and harms Jill while policy *B* benefits Jill and harms Ira, then there is no way for policy makers to judge which results in greater welfare unless they can compare how much Ira and Jill are each benefited and harmed by *A* and *B*. Since the benefits and burdens of alternative policies typically fall on different individuals, there is no way to compare the total or average welfare resulting from alternative policies without making interpersonal comparisons. Many nonutilitarian ethical systems also require interpersonal comparisons of well-being. To be rationally benevolent, one must be able to judge where one's efforts will do the most good. To treat the interests of different people equally, one must be able to compare the net effects of one's actions on the interests of each.

Interpersonal comparisons are also crucial to a classic economic argument supporting the redistribution of wealth and income. Although utilitarians are indifferent to the distribution of *welfare*, they are not indifferent to the distribution of income, because the amount of welfare that results from a given amount of income depends on how it is distributed. Economists writing near the beginning of the last century, such as A. C. Pigou (1920), argued that total welfare is maximized by equalizing incomes as much as is consistent with retaining incentives to produce. Citing the diminishing marginal utility of income, they maintained that, for example, an extra thousand dollars contributes less to the well-being of someone with an income of fifty thousand dollars than to the well-being of someone with an income of ten thousand dollars. Other things being equal, then, a more equal distribution of income increases total welfare. This argument assumes that one can make interpersonal comparisons of the amount that a thousand dollars contributes to the well-being of different people with different incomes. If interpersonal comparisons cannot be made, then this argument cannot be made either. So-called unit comparisons of utility *differences* are needed to compare the benefits and harms policies might cause to different people. One need only compare how individual utilities change, not their absolute levels. Comparisons of utility levels are needed if, for example, policy makers want to know who is worst off.

There are obvious difficulties in comparing how well off different people are. The main way economists and philosophers have attempted to understand interpersonal comparisons is via what Kenneth Arrow (1978) calls "judgments of extended sympathy" (see also Kolm 1972). Suppose we ask people to express preferences not only among ordinary alternatives but also among "extended" alternatives, such as the alternative of being Ira with

some option *x* and being Jill with *y*. We might then say that Jill-with-*y* is better off than Ira-with-*x* if people prefer to be Jill-with-*y* to being Ira-with-*x*.

This way of understanding interpersonal comparisons faces serious problems. It passes the buck, since the extended preferences agents have would seem to depend on interpersonal comparisons. Moreover, if people disagree about the ranking of being Ira with *x* and being Jill with *y*, *whose* extended preferences should decide? And more precisely, what exactly is the connection between, on the one hand, whether Ira is better off with *x* than Jill is with *y* and, on the other, McPherson's or Satz's or Hausman's preferences between the extended alternatives? John Harsanyi (1977b) suggests that, rather than employing one's own preferences to compare Jill-with-*y* and Ira-with-*x*, one should compare how well off one would be with *x if one had Ira's preferences* to how well off one would be with *y* if one had Jill's preferences. (How this is supposed to be easier than comparing how well off Ira is with *x* and Jill is with *y* is not obvious.) In Harsanyi's view, there is a single impersonal extended preference ranking to which our empathic abilities are a useful although imperfect guide. Judgments resulting from putting yourself in someone else's shoes in this way can be used to construct interpersonal comparisons. Alfred MacKay (1986) calls this the "mental shoehorn" tactic (see also Griffin 1986, ch. 7).

It is doubtful that extended preferences can provide a basis for interpersonal comparisons. Even if these preferences were unanimous, as is often not the case, they answer the wrong question. Everyone might prefer to be Jill with *y* – even though Ira is better off – simply because they admire Jill more than Ira. One might, for example, prefer to be Keats (who died of consumption at age twenty-five) rather than Queen Victoria (who lived to eighty-one and occupied the throne for sixty-three years) yet nevertheless believe that Victoria enjoyed a higher level of well-being than Keats.

Most economists take the problems of making interpersonal comparisons as a decisive reason to reject utilitarianism. But, however problematic in theory, interpersonal comparisons seem entirely feasible in practice: people make them all the time. For example, as a rough approximation, policy makers can suppose that those whose situations are similar in terms of income, health, and personal relations are equally well off, or that those who have a very low level of income, health, and personal relations are worse off than those who have a very high level of income, health, and personal relations.

Utilitarianism remains an extremely influential ethical view, sometimes named explicitly and sometimes not, particularly when one is concerned with issues of public policy, because it makes ethical questions in principle matters of straightforward calculation. Such calculations may be difficult to

carry out, and in some cases, utilitarianism will not give any definite advice owing to problems in learning the effects of policies and in measuring their welfare consequences. But there will be clear cases, and the reasons for indecision and disagreement can be stated clearly.

7.3 Justifying Utilitarianism

Utilitarianism is a tempting ethical theory. In many cases it matches common intuitions; and it is plausible to think that morality is centrally concerned with human well-being and that the consequences of our actions matter to their assessment. Furthermore, utlitarianism offers guidance in those cases where people's intuitions conflict: it shows how even difficult moral questions can be decided. Which policy or action a person ought to adopt depends on the consequences of alternative policies. Utilitarians can even cite the imperfections of human knowledge of consequences to explain why ethical questions are so hard to answer. If actual ethical systems are to a considerable extent implicitly utilitarian, then the utilitarian can offer a plausible explanation for why moral codes differ in different societies. Leaving one's grandparents out to die in the cold may have been morally permissible among those living in the harsh conditions of the Arctic; it may have maximized welfare in those conditions. The same policy is morally impermissible in affluent societies, where it does not maximize welfare. According to utilitarianism, what is morally right or wrong depends on the consequences, which in turn depend on the facts of the case. Utilitarianism is "absolutist" in one sense – whatever action maximizes welfare is the morally right action to perform – but it is not absolutist in the sense of supporting inflexible formulas for conduct. Even something as apparently heinous as intentionally killing civilians in wartime might in some circumstances produce more welfare than refraining from doing so and thus turn out to be morally justified.

Ultimate questions of justification are very difficult, but utilitarians have two ways to defend their doctrine. The first justification rests on the claim that welfare or well-being is the only intrinsically good thing. If the basic ethical obligation is to promote good states of affairs, then utilitarianism will follow; this is the justification for utilitarianism offered by Mill and Sidgwick. A second justification relies on the notion of equal respect (Griffin 1986, ch. 9). Interpreting equal respect as giving equal weight to everyone's *interests* leads naturally to utilitarianism.

The defense of utilitarianism in terms of equal respect can alternatively invoke the idea of a hypothetical agreement on the part of rational agents

concerned to advance their interests. From an impartial point of view, one can argue that these agents will endorse whatever moral principle serves individual interests impartially. John Harsanyi argues that that moral principle is average utilitarianism (1955). Suppose one models an impartial moral perspective as the perspective of an agent who thinks that it is equally probable that he or she could be any of the n members of society. The expected well-being of such an agent will be $(1/n)U_1 + (1/n)U_2 + \ldots + (1/n)U_n$, where U_i is the well-being of the i'th member of society. $(1/n)U_1 + (1/n)U_2 + \ldots + (1/n)U_n$ is, of course, the average welfare in the society. So, Harsanyi argues, impartial agents concerned to advance their interests – that is, to maximize their expected welfare – will endorse average utilitarianism.

These two justifications point toward differing interpretations of utilitarianism. The first argument – in which utilitarianism derives from the attraction of maximizing the good – leads naturally to maximizing the total utility of all sentient beings (whether human or not), whereas equal respect or contractualist arguments point toward maximizing the average utility only of those sentient beings who are rational. As noted in Section 7.2, these two outlooks have sharply different implications with regard to population policy.

7.4 Contemporary Consequentialism

During the decades that preceded the 1970s, utilitarianism appeared to be almost dead. Although it continued to influence policy makers, most philosophers did not take it seriously as a moral philosophy. Most economists had earlier abandoned utilitarianism in the face of the difficulties posed by interpersonal utility comparisons. And to pound the final nails in the proverbial coffin, John Rawls, in *A Theory of Justice* (1971), offered systematic criticism as well as an alternative theory that was suitable for guiding policy. The resurgence in practical moral philosophizing that was so prominent in the 1970s usually took for granted some sort of reciprocity or rights perspective rather than any sort of consequentialism.

Yet by the end of the 1980s utilitarianism and consequentialism were again highly influential in both theoretical and applied moral philosophy. The consequentialists of that era defend very different ethical theories, with James Griffin, for example, developing a sophisticated objective-list variant of utilitarianism and Amartya Sen developing a nonutilitarian view of consequential evaluation in which rights, capabilities, and functionings play a more central part than well-being. For instance, Sen (1979) suggests

that consequentialists can regard rights violations as themselves bad consequences in addition to any welfare losses that may accompany them.

This work shares many of the features of modern economic theory and in many cases shows the influence of economic modeling. Consider that most of these consequentialists link ethics to the theory of rationality. John Harsanyi, for example, writes, "Ethics … is a theory of rational behaviour in the service of the common interests of society as a whole" (1977a, p. 43). Most moral theorists do not go this far; but, as Samuel Scheffler (1982) especially has stressed, refusing to make trade-offs among different objectives and to maximize some objective opens one to charges of irrationality. Consequentialist theorizing – with its close association between rationality and ethics – is thus more congenial to economists and more easily integrated into normative economics than are, for example, rights-based ethical views that are not easily opened to calculation and trade-offs.

The past generation of utilitarians and consequentialists has been influenced by developments in economics and game theory. For example, Matthew Adler's *Well-Being and Fair Distribution* (2012) owes as much to economists' discussions of social welfare functions as it does to philosophical discussions of justice. In Russell Hardin's utilitarianism, human ignorance of consequences and the difficulties of measuring and comparing utilities hold center stage, and concepts from game theory are put to work generating the outlines of a theory of property rights and its limits (Hardin 1988).

Of particular interest has been the development of consequentialist moral theories in which the valuable consequences to be maximized are things *other* than well-being (see especially Parfit 1984). One important value in many of these theories is the satisfaction of needs; even utilitarians emphasize it, though not as a fundamental and intrinsic good. For example, although James Griffin is sympathetic to "informed preference" utilitarianism, he argues that policy should focus on needs because a government can more easily determine what people need than what will satisfy their informed preferences (1986, ch. 3). Griffin's emphasis on the empirical tractability of needs is ironic, given how averse economists have been to distinguishing needs from mere preferences. This aversion has arisen not from empirical difficulties but rather from theoretical objections to drawing the distinction. (Here is a case when moral philosophers may be more practical than economists!) In political discussions of economic policy, concern about human needs is already ubiquitous, and if philosophers can provide both a rationale for taking needs seriously in social decision making and a principled way of drawing the distinction between needs and wants

(Braybrooke 1987; Thomson 1987; Dasgupta 1995; Sen and Nussbaum 1993; Nussbaum 2000), then economists can put their modeling tools to work to help devise policies that will satisfy needs.

A good example lies in the area of medical research, where difficult choices must always be made about how much to invest in searching for treatments for various diseases. Rational allocation of scarce resources across different fields of research requires assessing the costs of research, the probabilities of success, and the relative urgency of progress in different fields. It is plausible that philosophers can contribute to making such choices well by developing defensible judgments about the value of health and the relative importance of the needs that would be met by successful treatment of different diseases.

The deep problems of utilitarianism – in particular, those concerning interpersonal comparisons of well-being – do not preclude operationalizing utilitarianism by providing a specification of the utility function to be maximized. One could, for example, stipulate a single utility function that roughly represents everyone's preferences. Preferences represented by such a common utility function would most plausibly be defined not over all marketed goods and services but instead over fundamental goods such as nutrition, shelter, or clothing that the agent extracts or constructs from marketed goods and services such as apricots, apartments, and aprons (Becker 1976; Kolm 2002). Behind the large differences in people's manifest preferences there might be agreement in preferences among the underlying goods. For example, Jill may prefer to eat at home while Jack prefers restaurants. The differences in their preferences may result more from differences in the opportunity costs they face in producing good meals at home than from any differences in underlying preferences. It is not absurd to postulate (as an approximation) a common utility function, and in terms of that function there might then be little problem determining the total utility of alternative policies – apart from the general difficulties of predicting their consequences. But we are in no position to judge how useful the idea of a shared "deep" utility function underlying apparent differences in preferences may be.

7.5 Is Utilitarianism Plausible?

The most powerful objection to utilitarianism is that it clashes sharply with many of our moral intuitions. What if a hereditary caste society results in more total happiness than a liberal democracy? A utilitarian would have to opt for the hereditary caste society. Or suppose that false testimony by

witnesses to the recent killings in the United States of African Americans by police would have led to indictments and convictions and thereby have prevented some of the severe rioting these incidents have provoked. It would seem that a utilitarian should favor perjury in cases such as these. Even if the convictions were unjust, the overall consequences of perjury would appear to be far better than the consequences of telling the truth. Yet most people believe that perjury is morally impermissible.

There are four possible responses to this apparent conflict between utilitarianism and what most people believe is right and wrong. First, utilitarians can argue that the conflict derives from a misapplication of utilitarianism. As John Rawls insisted in his "Two Concepts of Rules" (1955), one should distinguish questions about the design of institutions from questions about enforcement of their norms. Even if perjury would in these cases maximize utility, utilitarians would not favor laws and customs permitting perjury whenever witnesses conscientiously believe that perjuring themselves would maximize utility. So law and custom should condemn perjury, even in a case when committing perjury might maximize utility. This is not an argument for rule utilitarianism. It establishes only the weaker conclusion that perjury should be illegal and socially proscribed. It might still be morally obligatory for witnesses to break the law if they can do so secretly.

Second, utilitarians may object that cases such as the last one presuppose knowledge that is unattainable (Hardin 1988). When one takes into account the unavoidable uncertainties, it may turn out that utilitarianism does not in fact recommend perjury. This way of reconciling utilitarianism and moral intuition is not entirely satisfactory, because many people would say that whether perjury is morally right or wrong in this case does not depend on whether it would be detected or on whether it would ensure a conviction.

A third response to the apparent conflict between utilitarianism and moral intuition is to challenge the authority of "intuition." A utilitarian favors educating people to have strong and unambiguous moral convictions that promote desirable conduct in typical situations. It will be good on the whole if people are strongly moved by such feelings, but their considered moral judgments have no independent evidential force that can help in resolving hard cases (see Hare 1981).

In reply, the critic of utilitarianism can question what basis for morality there could be apart from intuition and our considered reflection on it. If people cannot take the intuitions they retain in the face of serious reflection seriously (though not uncritically), then they have no foundation upon which to argue for or against moral principles. So one arrives at a fourth

response to the apparent conflict between intuition and utilitarianism, which is to reject utilitarianism.

7.6 Consequentialism and Deontology

It might be thought that the general structure of evaluation we have called consequentialism is not subject to similar intuitive objections. Because consequentialism does not specify what counts as good, one might think that all ethical theories can be regarded as consequentialist. For example, those moralities that stress freedom rather than welfare might be regarded in consequentialist terms as seeking to maximize freedom, while those that stress duties might mandate maximizing conformity with duty. But this defense of consequentialism misunderstands the structure of many moral theories, which do not solely aim at maximizing some final value, such as freedom or duty. Many moral theories also give to people another end for the sake of which the final goal, whether it be freedom or even welfare, is pursued: the persons about whom the agent cares (Anderson 1993). Even if lying to a person would promote his freedom, it is unacceptable to do so on these theories because it is a disrespectful way to treat him.

In Scheffler's terminology, deontological (nonconsequentialist) ethical theories employ both "agent-centered prerogatives" (they sometimes *permit* agents to act in a way that does not maximize the good) and "agent-centered constraints" (they sometimes *prohibit* agents from acting so as to maximize the good). Agent-centered prerogatives and constraints are puzzling. How can it be morally permissible, let alone morally obligatory, to choose the lesser good? Deontological theories not only conflict with consequentialism, they appear to conflict with rationality itself.

Consequentialists might claim that if a person is serious about regarding killing as wrong and if she is convinced that murdering one innocent person will prevent the murder of two others, then it is only irrational squeamishness rather than moral principle that prevents her from committing the single murder. But refusing to murder an innocent person does not appear to be mere squeamishness. Acknowledging this, the consequentialist might argue that cases in which one can, with certainty, prevent two murders by committing one almost never occur. Individuals are thus better off following rules that prevent the killing of the innocent.

Conflicts between everyday moral principles and maximizing overall welfare are real and not uncommon. Consider, for example, the case of Baby Jessica, a toddler who in 1987 fell into a well in Texas. After a massive human effort costing hundreds of thousands of dollars, she was rescued.

The resources that were used to rescue her could have prevented the death and injury of hundreds of other American children if they had been devoted instead to better prenatal care. They could have saved the lives of thousands of malnourished infants in less developed countries. The distribution of health-care resources is shot through with similar examples.

Consequentialism implies that resources should go where they do the most good. Nonconsequentialists would counter that we owe more to someone at immediate risk of death than we do to those who face small risks, even if failing to attend to the small risks leads to a larger number of deaths (Kamm 1993). It is worth underscoring that deontological constraints need not be absolute. Whereas it might be morally wrong to kill one person to save two lives, it might be morally permissible to shoot down a hijacked airliner heading for a nuclear power plant that would kill hundreds of thousands. Yet, even with such a qualification, the consequentialist rejects deontological constraints as irrational.

In theory, it is possible to bring about a reconciliation between rationality and absolutist or deontological restrictions on intentional killing by distinguishing between *x being murdered* by someone and *my murdering x*. If that is the case, then there is no inconsistency in my preferring two others being murdered by some other person to my murdering *x*. Similarly, there is no inconsistency in preferring the death of a single child, Baby Jessica, to the death of hundreds yet preferring the death of hundreds of children to my standing by and letting Baby Jessica die. In order to make such seeming reconciliations of these two different moral perspectives more than ad hoc gimmicks, defenders of consequentialism must say something about why it is rational to make these distinctions and put so much weight on them. Here again we can see that utility theory presupposes a substantive background theory of "rational requirements of indifference" specifying when it is rationally permissible to distinguish among outcomes (cf. Section 4.3). But there is a new wrinkle. In this case, the background theory, which here determines when it is rational to distinguish among actions and outcomes, depends on substantive moral principles.

One proposed reconciliation allows the assessment of the goodness or badness of a set of consequences to vary with the perspective of the person doing the evaluating (Sen 1982b). An onlooker might conclude that it is morally better for one person to be murdered rather than for two to be murdered or for one child to perish rather than hundreds. But from the standpoint of the decision maker who is the acting party, what is at stake is whether *I* commit a murder or whether *I* instead stand by and let two children die by other processes. Everyone might agree both that two murders

are worse than one and that, from the point of view of an agent facing the prospect of committing a murder, carrying out the murder is morally worse than failing to prevent two murders from happening.

It might seem that evaluator relativity is less likely to be a consideration in public policy than in personal morality because policy should be made from an impersonal point of view. But as the Baby Jessica case and other examples from health-care policy show, it is not obvious that policy should attempt simply to maximize overall well-being. If Americans were consequentialists and did not rescue the likes of Baby Jessica – but also did not look the other way when hundreds of thousands of other children needlessly died – then the United States would in important ways be a better society. But if Americans could ignore Baby Jessica's plight with no more than a self-congratulatory recognition of how rational they were, who knows how inhumane they might become?

7.7 Conclusion: Should Economists Embrace Utilitarianism?

Welfare economists will find some versions of utilitarianism attractive because, apart from relying on interpersonal utility comparisons, they are so similar to standard welfare economics. Just define the rough-and-ready utility functions that will represent the preferences of the individuals affected and stipulate a way of making the interpersonal comparisons, and the way to utilitarian policy analysis is open. Welfare economists would not have to give up their focus on outcomes or their view of welfare. Of course, it is a bold and controversial step to stipulate what the utility functions should be and how the utilities of different people should be compared, but this step is no bolder than the identification of welfare with the satisfaction of preferences. Taking this further step would enrich normative economics.

Yet many of those concerned with policy would take issue with the suggestion that normative economists should become more explicitly utilitarian, precisely because they would like to see economists make a more radical break with current practice. Before considering this critique, let us turn in the next two chapters to a more detailed treatment of current practice.

Suggestions for Further Reading

Classic writings on utilitarianism include Bentham (1789), Mill (1863), and Sidgwick (1901). The collection of essays by Plamenatz (1967) provides a number of helpful philosophical treatments of problems of utilitarianism.

An old but still useful exchange on utilitarianism is Smart and Williams (1973). A helpful volume with emphasis on economic aspects is Sen and Williams (1982).

The most important consequentialist works are those of Brandt (1979), Broome (1991b, 2004), Griffin (1986, 1996), Hardin (1988), Hare (1981), Harsanyi (1977a), Kagan (1989, 1997), Parfit (1984, 2011) Railton (1984), Sen (1982b), and Singer (1979). For discussions of rule utilitarianism see Hooker et al. (2000), Hooker (2001), and Parfit (2011).

Applications of consequentialist moral philosophy to practical issues abound. See particularly Glover (1990) and Singer (1986, 2002). For discussion of the rationality of deontological moral principles that permit options not to do what maximizes the good or forbid maximizing the good, see Nagel (1986, ch. 9), Scheffler (1982, 1988), Kagan (1989), and Kamm (1993).

Questions for Study and Discussion

1. Utilitarian judgments of outcomes and policies are always comparative. Why? Is the fact that utilitarianism makes no absolute judgments a virtue of the view or an objectionable feature?
2. Do you think that utilitarianism is a reasonable guide to how we should treat nonhuman animals?
3. Do you think that a stronger case can be made for total utilitarianism or average utilitarianism?
4. It is often said that utilitarianism provides a plausible criterion concerning what is right or wrong but not a plausible method of making moral decisions. Why? What is the difference?
5. What do you think of Harsanyi's argument for utilitarianism in Section 7.3? Are you convinced? Why or why not?
6. Utilitarianism is an objective theory of what is right and wrong. If there is no alternative to *P* that leads to more total welfare, then *P* is morally permissible, regardless of what anybody in any society thinks. Yet, as the example of abandoning the elderly to freeze or starve illustrates, utilitarianism is open to the possibility that what is right in one society may be wrong in another. How is this possible?
7. Why are not all moral theories consequentialist? Are nonconsequentialist moral theories irrational?

8. In your opinion would it have been better if the resources devoted to saving Baby Jessica had been used instead to save many lives in less developed countries?
9. Suppose some people are much better at converting dollars into satisfactions than others. Maximizing society's utility would require giving these people disproportionately larger amounts of income than others. Is this fair? Is it right?
10. As we noted in Chapter 1, inequalities in the distribution of wealth and income in the United States are large and growing rapidly. According to Saez and Zucman, "The 16,000 families making up the richest 0.01%, with an average net worth of $371 million, now control 11.2% of total wealth – back to the 1916 share, which is the highest on record." Do you think that utilitarians should favor policies to lessen inequalities in income and wealth in the United States today? Why or why not?

EIGHT

Welfare

When people in modern Western cultures think about morality, they tend to focus on what is morally permissible or impermissible, right or wrong. But there are other important matters of moral concern: questions about what is good or bad and, more specifically, about what is good or bad for people.

What is good for a particular agent, Scrooge, will depend on factors such as Scrooge's character, ability, and circumstances, and what is good for Scrooge may differ from what is good for Marley. Most of the differences will likely concern *instrumental* goods – things that are good because they are means to something else that is good. If one focuses on *intrinsic* goods – things that are good in themselves, independently of their consequences – then there will tend to be less variation from individual to individual. Size 7 shoes are good for Scrooge while size 12 shoes are good for Marley, but both pairs serve the same end of walking. If nothing were good in itself, then nothing could be good as a means to some other end. There must be intrinsic goods in order for there to be instrumental goods.

In addition to its concern with right and wrong, moral philosophy offers theories about which things are intrinsically good for human beings. All plausible moral theories assign an important place to conceptions of such a good. This is obviously true of utilitarianism. However, even nonutilitarian views that emphasize rights, fairness, and justice need a conception of human well-being, if for no other reason than that these theories recognize the importance of benevolence, which requires some view about what makes people better off. Even the core notions of nonutilitarian theories often make reference to well-being. For example, most theories of justice are concerned with how well individuals' basic interests are met, and theories of right action often involve avoiding harm to people. Notions of harms and interests are plainly connected to notions of well-being.

8.1 Theories of Well-Being

In chatting with one's neighbors, as in studying moral philosophy, one finds many different theories of well-being. According to some religious views, the ultimate good lies in a specific relationship with God, while in others one's relationship with other humans is first and foremost. Some philosophers believe that only mental states are intrinsically good, but there is less agreement here than it seems because there are so many different views concerning *which* mental states are intrinsically good. Jeremy Bentham holds that the only intrinsic good is pleasure, while John Stuart Mill holds that it consists of "happiness," which involves judgments and satisfactions of a higher quality than the pleasures available to nonhuman animals. Mystics find the good in contemplative states of mind. Henry Sidgwick argued for a hybrid view that the good is any mental state (such as happiness, pleasure, hope, or love) that is intrinsically desirable.

Other philosophers endorse as intrinsic goods a potpourri of nonmental states ranging from human health and intimate relationships to achievements. The theory of well-being is a messy area of philosophy, and all the various philosophical theories face serious difficulties. These controversies are enough to send economists running back to their graphs. But economists cannot avoid thinking about well-being if they want to be able to judge when welfare increases or decreases.

Theories of well-being or welfare can be classified as formal or substantive. A substantive theory of well-being explicitly delineates what things are intrinsically good for people. Hedonism is an example of a substantive theory of well-being; its adherents hold that well-being is pleasure. Formal theories of well-being specify how to determine what things are intrinsically good for people, without identifying what those things are. To maintain that welfare is the satisfaction of preferences is to offer a formal theory of well-being. This theory does not entail anything about what things are good for individuals. A formal theory can be compatible with a substantive theory. For example, if happiness is the ultimate object of preference, then it is both the case that well-being is the satisfaction of preference and that well-being is happiness.

8.2 Welfare in Economics

Most economists do not directly address questions concerning welfare. Among those economists who do write about welfare, most are attracted to a formal theory, no doubt because formal theories appear to involve less controversial philosophical commitments than substantive theories.

By leaving the substantive question of what is good up to the individual, normative economists aim to show their philosophical modesty. Measuring well-being by how well satisfied an agent's preferences are also appeals to the antipaternalist view that individuals are the best judges of their own well-being. But, as we shall argue in this chapter, welfare is not best understood in terms of preference satisfaction.

Given mainstream economists' commitments to utility theory in explaining human choices, it is natural that they would look to levels of utility – that is, to the extent to which states of affairs conform to an agent's preferences – as the fundamental measure of human well-being for evaluative purposes as well. If individuals are exclusively self-interested, then they will prefer *x* to *y* if and only if they believe that *x* is better for them than *y* is. If they are perfectly well informed concerning the facts and are good judges, then it seems that what they prefer will coincide with what is better for them. It follows that if a person, Jill, is a competent evaluator, self-interested, and well informed, then Jill prefers *x* to *y* if and only if *x* is actually better for her than *y*. Assuming that individuals are generally rational and competent judges and that they are self-interested and well informed, economists find it natural to identify welfare and preference satisfaction. In their applied work, economists often rely on more objective measures of "real income" rather than utility measures, but most economists view this as a compromise with data limitations. They regard real income as an imperfect proxy for preference satisfaction.

There are two possible explanations for the coincidence of preference satisfaction and welfare. One possibility, which we call "the constitutive view," is that Jill's preferring *x* to *y* when Jill is self-interested and Jill's factual beliefs are correct *makes it the case* that *x* is better for her than *y*. On this interpretation, normative economists are committed to the philosophical thesis that well-being *consists* in the satisfaction of well-informed self-interested preferences. Although what people actually prefer may not be good for them because their preferences are not directed toward themselves or because they hold false beliefs, actual preferences will typically coincide with well-informed self-interested preferences and will determine what is good for individuals.

We call the second explanation of why the satisfaction of self-interested and well-informed preferences is a guide to well-being "the evidential view." It is much more modest philosophically. If economists assume that agents are *competent evaluators* – that is, that when their beliefs about the consequences and properties of alternatives are true, they are generally good judges

of what is good for them – then when Jill's beliefs are true, Jill's judgment that *x* is better for her than is *y* is reliable evidence that *x* is in fact better for her than is *y*. On the evidential view, economists need not accept any substantive theory of well-being. Nevertheless, in order for economists to figure out when the evidence is flawed and when it is accurate, the notion of well-being cannot be completely empty. If economists suppose that people are generally self-interested and evaluatively competent, how do they determine which are the contexts in which this general supposition is unjustified? We think that platitudes that are part of the everyday view of well-being provide some guidance. In those contexts in which people prefer the sorts of things that we think are generally better for people, such as greater health, closer friends, greater financial security, and so forth, economists have grounds to believe that people are more or less self-interested and competent evaluators. In contexts where people's preferences do not generally conform to platitudes such as these, one should be cautious of drawing any inferences concerning well-being from people's preferences. There is no need to endorse any special philosophical theory in addition.

Both the constitutive view and the evidential view justify measuring well-being by the satisfaction of preferences, although both recognize that what people prefer and what is good for them will differ when their preferences are not self-interested or when they depend on false beliefs. Welfare economists seldom mention these two important qualifications, and in practice economists sometimes operate with the mistaken view that everyone is self-interested. But, of course, people are not always self-interested. Only sociopaths are unable to distinguish what they believe to be best, all things considered, and what they believe to be best only for themselves. For most people, the two do not always coincide. Someone who is taking steps to prevent environmental disaster two centuries from now is not satisfying her self-interested preferences. Nor do people always have good information. Satisfying people's false beliefs about the properties and consequences of alternatives is not necessarily best for them. Drinking poisoned water that one believes to be safe does not make one better off.

There are circumstances in which it is a reasonable approximation to suppose that people are competent evaluators and that they are self-interested and have true beliefs, and in those circumstances, both the constitutive and the evidential view imply that well-being coincides with the satisfaction of people's actual preferences. One reason why the coincidence of well-being with preference satisfaction appeals to economists is that most economists dislike paternalism – that is, coercion that aims to benefit the person who is coerced. If what people want coincides with what is good for them, then

no coercive interference with their actions could possibly benefit them. However, this is not a good reason to take welfare to coincide with the satisfaction of actual preferences. Paternalism is sometimes desirable. Grabbing someone to prevent her from stepping into an open manhole that she did not see is one simple example of a justified paternalistic action; so arguably are seatbelt laws.

There are serious objections to the view that one can measure welfare by an agent's self-interested preference satisfaction, whether that preference satisfaction is interpreted as evidentially or as constitutively relevant to well-being. Two are particularly obvious. The first is that, unlike the purely self-interested creatures who populate a good deal of economic theory, people care about more than their own well-being. People sometimes sacrifice their own well-being in order to benefit others they love or to do harm to those they hate. And while almost no one's welfare is affected by whether the continuum hypothesis turns out to be true, it is easy to see how mathematicians could have a preference about that. Many people prefer that their unborn great great grandchildren have good lives, even though that will have no bearing on their own well-being.

The second objection arises from the fact that people are ignorant of many things. Consequentially, people may prefer something that is bad for them because they mistakenly believe it is beneficial. It is not true that x is better for A than y if and only if A prefers x to y. Indeed, many people live in circumstances in which their governments, their poverty, or their lack of education makes it almost impossible for them to make informed choices.

8.3 Against the Constitutive View: Well-Being Is Not Preference Satisfaction

It is not clear whether economists are committed to the constitutive view of the relationship between preference satisfaction and well-being. It is more charitable to interpret the connection between preference satisfaction and well-being that they rely on as reflecting only an evidential connection. The evidential view is not without its problems, but these problems are small compared to the philosophical difficulties attaching to the constitutive view.

Before presenting these difficulties, we should point out that our rejection of the view that well-being is the satisfaction of informed and self-interested preferences is opposed to the philosophical consensus, which judges these theories to be among the most promising of theories of well-being. The views of philosophers are, however, somewhat different from those of economists. While economists (on our interpretation) leave

implicit the restriction to self-regarding preferences that are undistorted by false beliefs, philosophers such as Gauthier (1986, ch. 2); Goodin (1986); and Griffin (1986) develop with care restrictions on which preferences should be considered. They defend a theory of well-being as the satisfaction of well-informed, self-regarding, and "rational" preferences. Notice that taking well-being to be the satisfaction of rational preferences rather than actual preferences free of mistaken beliefs shifts the emphasis from what people in fact prefer to what is rational for them to prefer. If it could be shown, for example, that it is rational for everyone to prefer happiness to unhappiness or virtue to vice, then a view of well-being as the satisfaction of rational preferences would lead to a substantive view of well-being as involving happiness or virtue.

A number of considerations have made a constitutive theory of well being as preference satisfaction seem attractive. It offers an immediate explanation for the fact that different things contribute to the well-being of different people. It can be deployed without taking sides in the different substantive ends that people pursue. Indeed, among the many things that apparently make life go well are successes in the worthwhile projects that individuals have chosen. It is only because individuals prefer to pursue those projects that those successes contribute to their lives. If Henry V did not actually want to vanquish the French at Agincourt, then it would have meant little for him to succeed at it. Moreover, because what satisfies people's preferences (at least if they are well informed and self-interested) typically makes them happy, the preference satisfaction view of well-being can explain why hedonism (the view that well-being is pleasure) has been so influential.

In addition, because well-being seems to involve satisfaction, it might seem plausible to associate well-being with preference satisfaction. But this thought rests on an equivocation on the word "satisfaction." The satisfaction of preferences is like the satisfaction of degree requirements. It has no necessary connection to any *feelings* of satisfaction. To satisfy a preference for state of affairs x over y is for x to obtain rather than y. *Knowing that* x obtains may lead one to feel satisfied. But otherwise there is only a contingent connection between the satisfaction of a preference and any feeling of satisfaction. Satisfying the preferences of those who, concerned about climate change, want the Earth to remain habitable three hundred years from now will give them no feeling of satisfaction, because they will be long dead.

Despite the considerations that may make preference satisfaction theories of well-being seem plausible, these theories should be rejected. Recall

that we have described them as maintaining that the satisfaction of well-informed and self-interested preferences constitutes well-being. As careful readers may already have noticed, what can it mean to say that some preferences are "self-interested"? In ordinary language, Satz's preferences are self-interested if and only if they are directed toward Satz's own well-being. To maintain that Satz's well-being consists in the satisfaction of those preferences of Satz that are directed toward Satz's well-being presupposes some other theory of well-being than preference satisfaction. Philosophers have, unsurprisingly, noticed the problem. At the same time, some restriction on preferences is needed, lest one is forced to conclude that the satisfaction of preferences that have nothing to do with people's own lives nevertheless makes them better off. Consider the following example of Derek Parfit's:

> Suppose that I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I strongly want this stranger to be cured. We never meet again. Later, unknown to me, this stranger is cured. On the Unrestricted Desire Fulfillment Theory, this event is good for me, and makes my life go better. This is not plausible. We should reject this theory. (1984: 494)

To restrict the class of preferences whose satisfaction constitutes well-being without surreptitiously invoking some other notion of well-being to distinguish self-interested preferences, philosophers have attempted to distinguish those preferences that are "about" the agent from those that are not. For example, Mark Overvold (1984) proposes that an agent's well-being should be regarded as the satisfaction of those among the agent's preferences that concern states of affairs that entail the agent's existence.

A different way to respond to the problem that the satisfaction of Parfit's preference that the stranger be cured seems not to make Parfit any better off is to conclude that whether or not Parfit prefers x to y has by itself no effect on whether x is better for Parfit than is y. Why should anyone have ever thought that some state of affairs x that would otherwise be no better for Jack than is some other state of affairs y becomes better for him simply because he comes to prefer x to y? Suppose Jack has a preference concerning some aspect of his body that is hard to observe and irrelevant to his health: perhaps he prefers to have a small benign tumor on his spleen. This is clearly a "self-regarding" preference; it concern states of affairs that entail his existence. Is it plausible to believe that, owing to his preferences, Jack's possessing a small benign tumor on his spleen makes him better off?

Defenders of a preference satisfaction view might respond that all of Jack's preferences must be based on some advantage he sees for himself. Unless that advantage depends on a false belief, it shows that what Jack

prefers is, after all, better for him. But such a defense implicitly denies that well-being is the satisfaction of preferences. If a small tumor on his spleen is better for Jack, it is because of the advantages that led him to form his preference, rather than the preference itself. There is a gap between what people prefer and what is actually good for them. It is important not to confuse the satisfaction of a preference with coming to know that a preference is satisfied. If Jack learns he has a benign spleen tumor, he may be pleased and in this way he is better off. But to point to people's pleasure at learning that their preferences are satisfied is to take well-being to be pleasure, not preference satisfaction.

The defender of the constitutive theory might reply that even if preferences cannot conjure value out of nowhere, they are necessary in order for an agent's activities to promote her well-being. Suppose that Jill is very serious about bowling. She works hard at improving her skills and concentration and her skills improve. Her improvement gives her pleasure, but the contribution that Jill's bowling makes to the value of her life cannot be captured by the pleasure she takes in the activity. Her life is better for the successes she achieves in this activity, and if it were not for her preference for this activity over others, she would not have had nearly so much pleasure, and the successes she achieves would be unlikely and would not contribute as much value to her life. Our response is that although it is crucial that Jill prefer bowling to other activities, what actually makes her better off is her achievement and the pleasure it yields, not the fact that it satisfies her preferences. If economists want to be philosophically modest, then they should not accept the constitutive view, which commits them to a controversial philosophical theory of what constitutes well-being.

So far, we have not mentioned the problems that arise when people prefer things that are, by some criteria, objectively bad for them. Some of these preferences may reflect poor information; it might be reasonable to believe that had the people had better information, they would not have prefered things that make their life go worse. But in other cases, lack of knowledge does not seem to be the relevant issue: would sadists who have access to full information about the harms they cause automatically change their preferences? This is doubtful.

The evidential view avoids these philosophical difficulties, and it serves the practical purposes of economists just as well as the constitutive view does. Measurements of well-being have many practical purposes: to establish benchmarks for equality, to assess progress over time, to compare living standards across communities, to weigh the comparative claims that different persons may justifiably make on social resources, and so on. What most

welfare economists care about is whether they can measure well-being by people's preferences, which the evidential view of the connection between preference satisfaction and well-being permits. In addition, as Sections 8.4 and 8.5 argue, the evidential view avoids a number of additional objections to accounts that define welfare in terms of preference satisfaction.

8.4 Conflicting Preferences and Well-Being

Measuring welfare by preference satisfaction leads to complications when preferences change. As Brandt (1979, ch. 13) and Parfit (1984, ch. 8) point out, if an individual's preference ranking changes, then it is unclear whether the individual is made better off by satisfying or frustrating the original preferences. Such cases seem to give rise to something akin to the problem of interpersonal utility comparisons: how does Jack's present self's satisfaction with *y* compare with his past self's satisfaction with *z* (Gibbard 1986)?

Should one care about satisfying preferences that people no longer have? Is there any reason now to satisfy Hausman's childhood desire to be a garbage collector? At this stage in his career, Hausman thinks not. Why not? Why should his current desires take priority over his past desires? One answer might be that his current desires not to be a garbage collector are stronger than his childhood desires. But as a six-year-old he longed *intensely* to drive a garbage truck. A second reason not to care about satisfying preferences that people no longer have is that giving people what they no longer want gives them no feeling of satisfaction. But this reason is not available on the constitutive view of the connection between welfare and preference satisfaction. In giving this reason one has shifted from a preference satisfaction theory to a mental-state theory of welfare.

The evidential view also enables us to give a third reason why satisfying Hausman's former preference will not benefit him: Hausman now has a much better understanding of what will make his life go well or badly than he had when he was six. He is (he hopes) a more competent evaluator. Of course, not all instances of preference change can be dealt with so easily. When people's preferences change, their later preferences do not always reflect better judgment than their earlier preferences.

This theoretical problem of preference change is linked to a practical problem: policies and institutions change people's preferences. (It is hard to believe that the billions spent on propaganda and advertising have *no* effect on preferences.) Assessments of policy must then depend in part on one's views concerning which institutions provide a suitable framework for developing desirable preferences (McPherson 1982, 1983b). Additionally,

how concerned should one be about satisfying current preferences if one judges that they are likely to change? Should one aim to modify preferences so that they will be easier to satisfy? How should one choose between either satisfying existing preferences or modifying preferences first and then satisfying the modified preferences?

On a constitutive view, these questions have no good answer. On an evidential view, the answer presupposes a substantive theory of well-being that enables economists to judge which preferences offer better or worse evidence concerning well-being. Economists do not have such a theory, and most of them do not want it. But, given preference change, there is a significant ambiguity about what satisfies preferences.

8.5 Assessing Preferences

On the constitutive view, provided that McPherson is well informed, one makes him better off by satisfying his self-regarding preferences, regardless of how idiosyncratic or obnoxious they are and regardless of how they were formed. But some of the real world uses of the theory of welfare demand that policy makers discriminate among preferences.

For example, as Thomas Scanlon has pointed out, both policy makers and public discourse rely on a relatively objective standard of "urgency" when weighing the strength of competing claims for social provision. The urgency of people's claims does not correspond to the strength of their preferences. For example, even if members of a destitute religious group prefer subsidies to build religious monuments over receiving food and shelter, their beneficent fellow human beings (whether fellow citizens or foreign donors) might acknowledge a moral obligation only to provide food and shelter (Scanlon 1975; see also Sagoff 1986). Without pretending to settle all questions about what makes people better off or worse off, and without taking any stand concerning what things are of ultimate value, a good deal of social policy rests on a settled agreement of what beneficence demands. What we owe to others is understood in terms of objective factors such as relieving hunger or homelessness, not in terms of the subjective preferences that agents happen to have.

One way of defending discrimination among preferences (which is compatible with an evidential but not a constitutive view of the connection between preferences and welfare) would be to argue that members of the religious group are mistaken about what promotes their well-being. An alternative interpretation – that perhaps accords better with the spirit of liberalism – is to insist that the policy decisions in a liberal state depend

on different standards than the decisions of individuals concerning what goods to pursue. The state's job is to provide for basic needs and to make available to individuals a wide range of activities, not to promote the subjectively valuable projects of individuals.

Here one can see, by the way, one reason for preferring "in-kind" provision to transfer payments. If donors provide cash benefits, destitute members of this religious group will not use the funds to alleviate the needs that gave rise to our obligation to assist them. Those who defend the view that policy should aim to promote welfare, where welfare is measured by preference satisfaction, might claim that our resistance to honoring idiosyncratic preferences is purely practical: we would open ourselves to manipulation and misrepresentation if we let people's reports of their subjective needs govern public distribution of benefits. One can also argue that the state has no business partnering with individuals in their own projects.

Should public benevolence or justice be sensitive to what people prefer? What moral pull should satisfying Hausman's preferences have on McPherson or Satz? Thomas Nagel has argued that if something is valuable to people only because they want it, then their getting it has no direct moral importance for others (1986, ch. 9). Others have no reason to satisfy Hausman's preferences unless they can make sense of why what he wants is worth wanting, or why his life will be better in some substantive way if he gets what he wants. On the constitutive view, there is nothing to be said about whether what people prefer is worth preferring, because to maintain that some things are not worth pursuing assumes that there is some source of value other than the satisfaction of preferences. On the evidential view, in contrast, there is nothing odd about asking whether something is worth pursuing – although it may be difficult to reach agreement about this in many cases.

Consider that many Boston residents desperately wanted the Red Sox to win the World Series in 2004. Their happiness when the Red Sox won gave others, even Yankees fans, some reason to want their preferences to be satisfied. But Nagel maintains – very plausibly, we believe – that the mere preferences of Red Sox fans (as opposed to their happiness or unhappiness or other grounds for their preferences) should be of no *moral* importance to others. This line of thought implies that social policy should not be concerned with satisfying preferences except insofar as doing so coincides, as the evidential view maintains, with other social objectives.

Consider those who have expensive tastes. An even-handed concern to satisfy their preferences appears to be unfair to those with more modest tastes. If welfare is preference satisfaction (or indeed happiness), then

a person who has cultivated a taste for "prephylloxera claret and plover's eggs" (Arrow 1973, p. 254) without an income that makes them affordable is worse off than someone with a similar income who wants only affordable beans and franks (see also Dworkin 1981a). But should social policy be responsive to expensive preferences? What is at issue may be fairness rather than well-being: the person with the taste for plover's eggs may well be worse off without the plover's eggs, but policy makers distributing scarce resources do not have to care.

Should the unsatisfied preferences of those with expensive tastes not count at all? Perhaps their unhappiness should count in some way, but why should their inability to procure the expensive things they want be a matter of social concern? Those who defend more objective views of well-being maintain that, except insofar as they are unhappy, the fancy eaters are not at all worse off. If one refuses to be disturbed about failures to satisfy expensive preferences, then either one's benevolence is limited or one does not accept the view that preference satisfaction constitutes well-being.

Racist, sadistic, and other antisocial preferences raise related problems for the constitutive view. Some of these may be based on false beliefs. But that is not always the case. One reason why it is bad to satisfy them is that doing so frustrates other (and characteristically stronger) preferences. Should policy makers balance the harms of these preferences against the "benefits" of satisfying them? Should antisocial preferences count at all (Harsanyi 1977a, p. 56)? Given either a hedonistic or a preference satisfaction theory of well-being, it seems that antisocial preferences ought to count and that a benevolent person should, other things being equal, strive to satisfy them. More objective views of well-being, on the other hand, such as the evidential view of the relations between well-being and preference satisfaction, permit one to deny that individuals are made better off when their antisocial preferences are satisfied. This is certainly a mark in their favor.

Defenders of a preference satisfaction view of welfare have some ways of deflecting the claims of those with antisocial and expensive preferences without questioning the identification of well-being and the satisfaction of preferences. First, they can claim that some additional moral consideration – such as justice, which is distinct from benevolence – is involved. Racist preferences do not count because they deny some people's rights. Second, they can note that preferences are malleable and that expressions of preferences respond to incentives. Frustrating expensive and antisocial preferences discourages their expression and may in the long run enable more preferences overall to be satisfied. These points are plausible, but they leave defenders of a constitutive view of the relationship between preference

satisfaction and well-being with the unsatisfactory implication that, other things being equal, there is reason to satisfy antisocial and expensive preferences – as if the extent to which racism limits well-being depends on how preference satisfactions turn out.

A final difficulty with the view that well-being coincides with preference satisfaction concerns preference formation. Some preferences derive from previous injustice, coercion, or manipulation, and people may form preferences as the result of problematic psychological mechanisms. Some people want things precisely because they cannot have them ("The grass is always greener on the other side of the fence"), while others spurn what is beyond their reach, like the fox who judged the unobtainable grapes to be sour (Elster 1983; Sen 1987b, 1990a). As Sen has pointed out, women who are systematically denied roles in public life or equal shares of consumption goods may learn not to want these things. Women who have never known freedom may not know how to value it. Women who have never been paid wages comparable to men's wages and have been subject to regular abuse may not demand fair wages or even physical security. But liberties, high wages, and protection from domestic violence will more plausibly make these women better-off than giving them what they may prefer. Satisfying preferences that result from coercion, manipulation, or "perverse" preference formation mechanisms does not always make people better off. Suspect preferences sometimes reflect false beliefs, and some of the problems here are versions of problems we have already discussed. But coercion and manipulation may also distort the psychological mechanisms that lead people to value some things more than others. For example, there is evidence that employers keep their bonded laborers separate from other wage workers so they cannot have access to information about leaving bondage; the employers seek to cultivate servility (Schaffner 1995). When the preferences of oppressed people derive from the circumstances of their oppression, one cannot measure their welfare by considering how well their preferences are satisfied. The constitutive view fails to register this point.

8.6 The New Hedonist Welfare Economics

Many nineteenth- and early twentieth-century economists were utilitarians who regarded well-being as a mental state such as pleasure or happiness. In contrast, modern economic theory, as developed in the 1930s, abandoned hedonism. Because economists found that the basic propositions of demand theory and consumer behavior require only that people have stable preference rankings with certain properties, most took well-being to be the

satisfaction of preferences. Even if many economists informally continue to associate welfare and happiness, standard welfare economics for the better part of a century has measured welfare by preference satisfaction.

Over the last two decades, a number of prominent economists, impressed with improved techniques for measuring subjective states, have adopted more sympathetic attitudes toward hedonism (Kahneman and Sugden 2005; Kahneman and Krueger 2006; Kahneman and Thaler 2006). They assume that people's preferences depend on their own estimates of the happiness alternatives will produce, and that direct measures of pain and pleasure would be more accurate measures of well-being. For example, Kahneman and Krueger (2006) define a misery index that they call the "U-index," which measures what proportion of their time people spend in unpleasant states. An espisode counts as unpleasant if the most intense feeling people report concerning it is negative. Kahneman and Krueger then measure people's subjective well-being by the percentage of their time that people spend in unpleasant episodes. To lessen the unhappiness they measure, Kahneman and Krueger suggest mental health interventions and policies to reduce the amount of time spent engaged in unpleasant activities such as commuting to work or doing housework. The proposal that welfare economics should focus on happiness ("experienced utility") rather than preferences ("decision utility") has already had a practical impact. Both the Organization for Economic Development and Cooperation) (OECD) and the British government have begun to collect systematic data concerning subjective well-being and have sought to appraise policies by their consequences for people's subjective states. The measures of subjective well-being rely on survey questions concerning people's affect, their "life evaluation," and their psychological flourishing. Something of the same ambiguity we saw between a constitutive and an evidential view of the relationship between preference satisfaction and well-being may arise here. It is ambiguous whether these measures assume that people's subjective states as measured by these surveys constitute their well-being or whether they provide evidence of their well-being, which need not be a subjective state.

Hedonism has obvious attractions. Pleasure strikes many people as intrinsically good, and pleasures seem to be measurable. Many different things give people pleasure, and so hedonism recognizes that different activities and outcomes are good for different people. But hedonism also faces serious objections. Pleasures and pains are diverse. The satisfaction of solving a math problem is a different kind of mental state from the delight of hearing some wonderful song or the comfort of a warm bath after a day outdoors in the cold. Unpleasant states are at least as diverse. How does one

put vertigo, nausea, and toothaches on a single scale with the pleasures of achievements, music, or baths?

Although contemporary psychologists have made progress on the measurement difficulties, there is a fundamental objection to hedonism and indeed to all mental state theories of well-being. Those who maintain that well-being consists in mental states are committed to the view that two individuals who are in the same mental state are equally well off, even if their objective circumstances differ. So someone whose life is going splendidly may be no better off than someone else whose life is going miserably but who mistakenly believes that it is going well. Robert Nozick (1974, p. 41) proposed a hypothetical case that illustrates this difficulty. Suppose there were an "experience machine" that could give people the highest quality experiences possible. These terrific experiences might be intense physical pleasures or they might be experiences of climbing Everest or composing a symphony that surpasses Beethoven's best efforts. (Once attached to the machine, people no longer know that they are attached; the machine takes over their consciousness and memories completely.) Suppose people attached to this machine experience whatever states of consciousness that mental-state theorists of well-being claim to constitute what is ultimately and intrinsically good. The mental-state theorist would then have to say that all people would be better off permanently hooked up to a reliable experience machine rather than living their own lives and experiencing the decidedly mixed mental states that will be part of any human life. Some people, whose lives are ones of intense and unremitting suffering, would arguably be better off on the machine. But if one believes that those who are hooked up to the experience machine are missing some of the intrinsically good things in life – even though they are not, by assumption, missing the best mental states – then one cannot accept a mental-state view of well-being.

We conclude that mental state theories of well-being, including hedonism, are mistaken. But just as preference satisfaction may provide evidence concerning well-being even if it does not constitute well-being, so subjective experience may indicate well-being without constituting it. The fact that hedonism is untenable as a theory of well-being does not rule out the proposal to measure well-being by people's subjective states.

Nevertheless, we are skeptical. We have already raised some doubts about whether subjective experience is a good indicator of well being. We also have doubts about the reliability of the measures of subjective experience. There is, as yet, no canonical way of measuring subjective experience, and the different measures give different results. Furthermore, when economists evaluate policies by subjective experience rather than by preference

satisfaction, they cannot piggyback on individuals' judgments of the ways in which policies bear on more distant ends. For example, some surveys of people's moods have shown that people find taking care of their children as unpleasant as housework (Kahneman et al. 2004). This finding does not justify policies designed to reduce the amount of time people spend rearing their children, because child rearing has many important and less immediate consequences. Standard welfare economics relies on the trade-offs people make between immediate affect and longer-term consequences, which are reflected in their preferences for child rearing over dishwashing. In contrast, economists who measure welfare by subjective experience need to rely on their own judgments of how the values of the longer-term consequences of policies should be balanced against their immediate pleasures or pains.

8.7 Other Theories of Well-Being

Substantive theories of welfare purport to say which things are intrinsically good. Hedonism is a substantive theory, as are "perfectionist" views (Griffin 1986, ch. 4; Raz 1984, ch. 12) and what Parfit calls "objective list" views. Substantive views are objective in the sense that what is good for people is not determined by whether people believe it is good for them. Even as they take well-being to coincide with the satisfaction of preference, many economists are also committed to a substantive view of well-being as material self-interest. These views are compatible if and only if people's preferences reflect their material self-interest.

As the example of hedonism shows, the objectivity of substantive views of well-being does not imply that subjective states do not matter, and most substantive theories give some weight to mental states. While objective views are controversial, they should nevertheless be tempting to economists insofar as they make well-being more readily measurable than are mental states.

In addition to material self-interest, one objective view of what matters for economic policy, which may be relevant to normative economists, is John Rawls's account of "primary goods." This is closer to a generalized notion of resources than to an alternative theory of well-being. In *A Theory of Justice* (1971) Rawls takes well-being to be the satisfaction of rational preference, but he does not think that justice should focus on well-being. One reason is that people's well-being depends in large part on their own efforts. Amartya Sen defends a theory of well-being in terms of the *capabilities* and the *functionings* an individual attains. A person's functionings consist in

those things that the person does and experiences. Walking, playing the piano, being well nourished, loving one's friends, understanding Chinese, and appreciating cubism are all functionings. But Sen does not think that social policy should concern itself directly with the functionings that people achieve. Social policy, insofar as it is motivated by a concern for welfare, should instead focus on capabilities (Sen 1987c, 1992a). A capability is the ability to achieve a certain sort of functioning. For example, literacy is a capability while reading is a functioning. People may value capabilities for their own sake as well as for the functionings they permit – you are glad to know you can walk around even if you are inclined to stay put.

In Sen's view, social policy should be less concerned with functionings – with what people make of their capabilities – than with capabilities, because functionings depend heavily on individual choices. For example, a shortfall in functioning such as malnutrition might stem from an individual's decision to embark on a religious fast rather than from any unjust deprivation. In contrast to Rawls, Sen maintains that social policy should not focus on resources or primary goods, because people with equal primary goods may lead very different lives owing to traits that are internal to the individual such as disabilities. For example, as a result of a digestive disorder, someone may be malnourished despite having a normal diet. In Sen's view, the focus of policy should be on ensuring that people have a decent set of capabilities, broadly understood.

Martha Nussbaum has also conceptualized human flourishing in terms of capabilities (2000, 2001). Her view of capabilities differs from Sen's in two ways. First, its roots are in Aristotle's moral philosophy rather than in the inadequacies of conventional welfare economics. Second, Nussbaum has formulated a specific list of central human capabilities.[1] She recognizes that some items on the list might be realized differently in different societies and that some items are more firmly fixed than others. Nonetheless, she argues that the list has survived considerable cross-cultural scrutiny. While Sen has also been concerned to make capabilities measurable and practical, he has been more cautious about specifying the most important human capabilities.

Nussbaum's and Sen's approaches complicate the measurement of well-being. People vary in the capabilities that are important to them, and they

[1] Nussbaum's list of central human capabilities (2001, pp. 416–418) consists of the following 10: (1) Life; (2) Bodily Health; (3) Bodily Integrity; (4) Senses, Imagination, and Thought; (5) Emotions; (6) Practical Reason; (7) Affiliation; (8) Other Species; (9) Play; and (10) Control over One's Environment, both Political and Material.

may disagree about the rankings of capabilities. Do preferences reenter to rank capabilities? Furthermore, figuring out what capabilities a person has is a fraught enterprise. Does Satz still have the capability to be a professional musician, or did she lose that when she gave up the piano at age thirteen? And did she have that capability at age thirteen?

Even if the problems of measuring the components of well-being and assigning weights to them prove to be greater for more objective approaches than are the problems of measurement in standard welfare economics, it can still be argued, following Sen (paraphrasing Carveth Read and John Maynard Keynes), that "it is better to be vaguely right than precisely wrong" (1987a, p. 6). In particular, objective approaches to well-being may lead research in directions that link up naturally to the normative terms characteristic of policy debate. For example, an objective index is much more helpful as a measure of the extent of deprivation in less developed countries than is utility theory. The index that is currently most used, the human development index (which is, roughly, an average of life expectancy, literacy, and per capita GDP), owes a great deal to Sen's views (and to Sen himself). A much more detailed set of indicators can be found in Anand et al. (2009).

8.8 Conclusions

Measuring well-being by the satisfaction of preferences is problematic. It mistakenly suggests that social policy should attend to all preferences – even if they are expensive, antisocial, or the results of false beliefs, manipulation, or problematic psychological processes. The focus on preference satisfaction does not cleanly link up with the normative terms of policy debate, and it leads to difficulties when preferences change and conflict. Yet as we shall see in the following chapter, economists have found ways of measuring preference satisfaction in order to inform policy, and, as we have noted, identifying well-being with the satisfaction of preferences ties together positive and normative economics. Whether the link is of value depends on how well economics succeeds in addressing problems of human welfare.

Suggestions for Further Reading

Defenses of an informed preference satisfaction view of individual well-being can be found in Gauthier (1986, ch. 2), Goodin (1986), Griffin (1986), and Arneson (1990). The introduction to *Utilitarianism and Beyond* (Sen and Williams 1982) summarizes arguments against preference-based

approaches. The central argument here derives from Kraut (2007) and Hausman (2012). See also Mill (1863), Rawls (1982), Sen et al. (1987), and Scanlon (1998).

Harris and Olewiler (1979), Hammond (1983), Machina (1987), Broome (1991b, ch. 10), and Hausman and McPherson (1994) discuss the problems that arise in respecting preferences when there are uncertainties.

For discussions of the new hedonist welfare economics see Kahneman (2000a, b), Kahneman et al. (2004), Kahneman and Sugden (2005), Kahneman and Kreuger (2006), Kahneman and Thaler (2006), Layard (2006), OECD (2013), and Vizard and Rusgys (2013).

Sen's views on capabilities and functionings are developed mainly in two of his essays (1985a and 1987c) and are discussed at length in Nussbaum and Sen (1993). Nussbaum (2000, 2001) attempts to recast political theory in terms of fundamental capabilities. For more on the human development index, see http://hdr.undp.org/en/content/human-development-index-hdi?

Questions for Study and Discussion

1. How does the theory that well-being is happiness differ from the theory that well-being is the satisfaction of preferences? Which theory do you think is more plausible?
2. If you were convinced that there were an absolutely reliable experience machine that would give you the subjective experience of what you believe would be an absolutely marvelous life for you, full of all the things that you value the most, would you choose to be connected? Why or why not?
3. What exactly is the difference between the evidential and the constitutive views of the relations between preferences and well-being? Under what circumstances are both of these views false? If there are many circumstances in which what is best for people does not coincide with what they want, is it not a mistake for economists to measure well-being by prefrence satisfaction?
4. What are the main objections to the constitutive view of the relationship between well-being and the satisfaction of preferences? Do you think that these objections are decisive?
5. It seems plausible to maintain that satisfying one of Jill's preferences makes her better off only if she has the preference at the time it is satisfied. Satisfying preferences that agents no longer have or do not yet have does not count. How would you explain this?

6. Some people have modest desires, while others are unsatisfied unless they have very expensive things. If public policy aims impartially to help people to satisfy their preferences, then those who want expensive things have greater claims on public resources than those who are more easily satisfied. Do you accept this implication? Why or why not?
7. What do you think makes for a good human life and what philosophical theory of well-being seems most plausible to you?
8. Do you think that welfare economics would be more useful if economists evaluated policies by the implications for happiness rather than by their implications for preference satisfaction? Why or why not?
9. Rawls's theory suggests that policies can be evaluated by their consequences for people's holdings of primary goods, while Sen and Nussbaum urge instead that policies be evaluated by their consequences for capabilities and functionings. What advantages do these proposals have over focusing on preference satisfaction, as economists do? What disadvantages?
10. Should policy makers count people's antisocial preferences in policy decisions? Why or why not?