Lecture Notes for *Ethics & Economics*

Draft of Spring, 2019

Contents

Chapter 1

Introduction to Ethics and Ethical Reasoning

Explain the normative concepts of goodness, rightness, welfare, and rationality, as we are using those terms in this class. What kinds of things are evaluable as good? What kinds of things are evaluable as right? What kinds of things are evaluable as good for you*? And what kinds of things are evaluable as* rational*? What is an ethical theory? (Provide some examples.) What makes an ethical theory* consequentialist *or* non-consequentialist*? (Provide an example of each.)*

1.1 Normative and Descriptive Claims

1. *Descriptive* claims say something about the way that the world is. They do not make any judgment about whether the way the world is is the way that it ought to be, or whether the way that the world is is a good way for it to be. For instance:

   (a) The Braves won the World Series in 1995.

   (b) It is always sunny in Philadelphia.

   (c) Even the poorest of the poor would have more wealth if we abolished the capital gains tax.

   (d) Johann believes that we should abolish the capital gains tax.

   Note that (b)—and maybe (c), though it's controversial—is false. So a claim need not be true in order to be descriptive. And note that, even though it *concerns* how the world ought to be, (c) still just describes a belief that Johann has; it doesn't endorse that belief, so it is merely describing the way that the world happens to be—it happens to be such that Johann believes that we should abolish the capital gains tax.

2. *Normative* claims say something about the way that the world *ought* to be, or something about which things are *good*, or which actions are *right*, or *rational*. They don't merely *describe* the world; they additionally *evaluate* the world. For instance:

   (a) We should abolish the capital gains tax.

   (b) We shouldn't abolish the capital gains tax.

(c) We have a moral obligation to ameliorate the suffering of the global poor.

(d) Johann shouldn't think that we should abolish the capital gains tax.

(e) If you want to steal a candy bar, you should go to the bodega on 5th avenue.

(f) Voting is irrational.

Note that either (a) or (b) is false; so a claim need not be true in order to be normative. Note also that a *normative* claim needn't be a claim about what's *moral*. At first blush, (e) gives an example of a normative claim which merely concerns what is *prudentially rational*, as opposed to what is *moral*—it says what you should do, in light of your desire to steal a candy bar; it doesn't say anything about what's *morally* best. Also, on the intended reading, (f) is not explicitly a claim about what's moral, but rather a claim about what's *prudentially rational*.

3. As we'll be understanding the term here, ETHICS is the systematic study of which normative claims are true and which are false.

   (a) Descriptive claims may nevertheless be *relevant* to normative claims. For instance, if it's true that even the poorest of the poor would have more weath if we abolished the capital gains tax, this is surely *relevant* to the question of whether we should abolish the capital gains tax. However, ETHICS is not primarily concerned with descriptive questions. It's goal is to answer normative questions.

   (b) Wait...what if normative claims *aren't* true or false? You caught me. I am presupposing a substantive view in the field of philosophy known as *metaethics*. (Metaethics is, roughly, the study of what normative claims mean, or what it is to accept a particular normative judgment.) In particular, I am presupposing a position known as *cognitivism*. The cognitivist says that normative propositions can be true or false. In fact, the majority of metaethicists nowadays are cognitivists, so this isn't all that controversial of a presupposition to make. In this class, we'll be assuming throughout that normative propositions can be true or false. That is, we'll assume that it makes sense to wonder whether claims like "we should abolish the capital gains tax", or "The more of your preferences are satisfied, the better things are going for you" are true.

1.2 Foci of Ethical Evaluation

1. When we evaluate things *morally*, we may evaluate:

   (a) *States of Affairs* as good or bad (or, in the comparative: better or worse).

      i. For instance, we might ask whether one distribution of goods amongst citizens is better or worse than another.

   (b) *Actions* as right or wrong

      i. For instance, we might ask whether it's wrong to institute a regressive tax policy, or whether capital punishment is right.

2. It is possible for your moral evaluations of states of affairs and actions to come apart.

   (a) Suppose that I kill one person to save the lives of five. You could think that a) the state of affairs of 1 person being killed and 5 people living is better than the state of affairs of 1 person living and 5 people dying, even while thinking that b) it was wrong to kill the one person.

3. However, your moral evaluations of states of affairs and actions *needn't* come apart in these ways.
4. When we evaluate things *prudentially*, we may evaluate:
   (a) *States of affairs* as good or bad *for a you* (or, in the comparative: better or worse for you).
      i. For instance, we may ask whether one distribution of goods amongst citizens is better or worse *for you* than another is.
   (b) *Actions* as rational or irrational.
      i. For instance, we might ask whether voting (or tax withholding) is rational or irrational.

1.3 Ethical Theories

1. An *ethical theory* is a general theory of some normative domain. It is an account which tells you, *e.g.*:
   (a) for every given state of affairs, whether that state of affairs is good or bad; or, for any two states of affairs, which is better.
   (b) for any given action, whether that action is right or wrong; or, for any two acts, which is *more* right.
   (c) for any state of affairs, whether that state of affairs is good *for you* or bad *for you*; or, for any two states of affairs, which is better *for you*.
   (d) for any given action, whether that action is *rational* or *irrational*; or, for any two acts, which is *more* rational.
2. Some terminology:
   (a) A theory of which states of affairs are good (better) is an *axiological* theory.
   (b) A theory of which acts are (more) right is a *deontological* theory.
   (c) A theory of which states of affairs are good (better) *for you* is a theory of *welfare* (or *well-being*)
   (d) A theory of which acts are (more) rational is a theory of *rationality* (or a theory of *rational choice*)
3. Contrast this with a weaker kind of ethical claim: an *ethical principle* is a claim that *some particular kind* of state/action is good/right/rational. For instance,
   (a) "any organism potentially possessing [a serious right to life] has a serious right to life even now, simply by virtue of that potentiality."[1]
   (b) "Adding additional people whose lives are worth living does not make things worse"
   (c) "Choosing an act which you know will make things worse than an alternative is irrational"
4. Finally, contrast both of these with a third kind of claim: a *particular ethical judgment*, like, *e.g.*,
   (a) Socialism is bad
   (b) Cultural appropriation is wrong
   (c) Voting is irrational

[1] from Michael Tooley, "Abortion and Infanticide". *Philosophy and Public Affairs*. 2(1). 1972. (Tooley doesn't accept this principle.)

5. When we engage in ethical deliberation, we will transition back and forth between ethical judgments, ethical principles, and ethical theories. If an ethical judgment conflicts with an ethical principle, or an ethical principle conflicts with an ethical theory, then we must make a decision about which to alter. Our goal is to bring our ethical theories, principles, and judgments into alignment with each other—to reach a *reflective equilibrium*.

1.4 Consequentialism & Non-consequentialism

1. Let's separate deontological theories out into two broad kinds: *consequentialist* theories, and *non-consequentialist* theories.

   (a) Roughly, *consequentialist* theories claim that whether an action is *right* is determined, in some way or other, by the *goodness* of certain states of affairs.

      i. The consequentialist thinks, that is, that the proper evaluations of actions is determined by the proper evaluations of *states of affairs*.

   (b) Roughly, *non-consequentialist* theories claim that whether an action is right is not just determined by the goodness of states of affairs.

      i. The non-consequentialist thinks that whether an act is right or wrong is not settled by the goodness of any states of affairs related to that act.

1.4.1 Versions of Consequentialism

1. The Act Consequentialist thinks that acts are to be evaluated by looking at the goodness of *their* consequences. Two flavors:

   (a) Actual Value Act Consequentialism holds that an act is right iff performing *that* act actually leads to a better state of affairs than any other available act—that is: iff performing that act *maximizes goodness*.

   (b) Expected Value Act Consequentialism holds that an act is right iff performing that act is expected to have better consequences than any other available act—that is: iff performing that act maximizes *expected* goodness.

      i. By this, we mean the *mathematical expectation* of the goodness of the consequences brought about by the act. (If you don't know what this means, don't worry—we'll learn more about this later on in the course).

2. Rule Consequentialism holds that an act is right iff it aligns with a *rule* which is such that, if everybody (tried to?) follow that rule, the consequences would be (expected to be?) best—that is: iff it aligns with the *rules* which maximize (expected) goodness.

   (a) Whereas actual and expected value consequentialism evaluates acts directly in terms of their actual or expected consequences, rule consequentialism first and foremost evaluates *rules* in terms of the consequences that would (likely?) result from everybody (attempting?) to follow them. Acts are then evaluated as right or wrong by looking at whether they are in line with the correct rules.

Chapter 2

Utilitarianism & Kant's Moral Theory

2.1 Utilitarianism

In class, we saw that Utilitarianism may be understood as the conjunction of three different normative claims about three different normative categories. Introduce and explain each normative category, and then say what the Utilitarian says about it. Finally, rehearse one objection to Utilitarianism.

2.1.1 Welfare and Welfarism

1. Recall: as we'll use the terms in this class, a person's *welfare*, or their *well-being*, refers to how good things are going *for that person*.

   (a) This isn't a moral concept; it is conceptually possible that things go well for an immoral person.

   (b) It is an individualistic concept—when we speak of *your* welfare, we are interested only in how things are going *for you*. Things could be going well for you, even if they are going terribly for many other people.

2. The philosophical position we will call *Welfarism* is an axiological thesis (a thesis about *goodness*) according to which what makes states-of-affairs good or bad is just the total amount of *welfare* that individuals possess in those states of affairs.

   Welfarism
   The goodness of a state-of-affairs is determined by the total amount of welfare that individuals possess in that state-of-affairs. A state-of-affairs, S, is better than another state-of-affairs, T, iff S has a greater amount of *net, aggregate* welfare than T does.

   (a) While *welfare* is an individualistic matter, the goodness of states of affairs is not, according to the welfarist. In order to see how good a state-of-affairs is, you must look at *everybody*'s welfare.

3. A philosophical theory of welfare will tell us *what it is* for a life to go well. It won't tell us what kinds of things *cause* a life to go well—maybe it's family, maybe it's wealth, maybe it's power, or maybe it depends upon the person, but a theory of welfare doesn't directly take a stand on such questions. Rather, it will say what family, wealth, or power would have to cause, in order to cause a person's

life to go well. (Compare: *what it is* for something to be hot is for its mean kinetic energy to be high. Saying this is not saying what *causes* things to be hot—the things that *cause* heat include stoves, fire, and sunlight. But our theory of *what heat is* won't mention these causes.)

(a) In brief, here are some sample theories of welfare:

i. What it is for your life to go well is for you to experience pleasure. (*Hedonism*)

ii. What it is for your life to go well is for you to get the things you desire. (*Preferentism*)

iii. What it is for your life to go well is for your life to be filled with knowledge, health, laughter, intimacy, and.... (*Objective List*)

iv. What it is for your life to go well is for your life to be spent on projects of positive value.

2.1.2 Utilitarianism

4. The *Utilitarian* accepts welfarism, and *in addition*, accepts a hedonic theory of welfare.[1] This supplies them with a full theory of the goodness of states-of-affairs (a complete *axiology*). The utilitarian then goes on to endorse *consequentialism*, which provides them with a deontological theory of the rightness of actions.

(a) That is, Utilitarianism is the conjunction of the following three theses: hedonism, welfarism, and consequentialism.

> Utilitarianism
> A Utilitarian is somebody who accepts each of the following claims.
> Hedonism: What it is for an individual to have greater welfare—*i.e.*, what it is for things to go well for an individual—is for them to have more pleasure and less pain.
> Welfarism: What it is for one state-of-affairs, S, to be better than another state-of-affairs, T, is just for S to have a greater amount of *net*, *aggregate*, welfare than T.
> Consequentialism: The rightness of an action is determined by the goodness of the states-of-affairs which result from the action's performance.

i. That is: utilitarianism is the conjunction of an ethical theory of welfare (hedonism), an ethical theory of goodness (welfarism), and an ethical theory of right action (consequentialism).

5. By substituting in different versions of consequentialism, we will get correspondingly different versions of utilitarianism. So, for instance, all of the following are versions of utilitarianism: act utilitarianism, rule utilitarianism, and expected value utilitarianism. Going forward, let's just focus on the expected value version of utilitarianism.

6. Telescoping their three theses down into one, and considering only what the expected value utilitarian has to say about right action, we get:

> Expected Value Utilitarianism
> An act is permissible iff no other available act leads to a higher level of **net aggregate expected pleasure**.

[1] The definition of utilitarianism I am providing here is a bit idiosyncratic and a bit anachronistic. I'm presenting things this way for pedogogical reasons. As we'll see later on, the ethical theory of welfare economics will fit into this schema, with preferentism exchanged for hedonism.

7. We should explain each of the bolded terms in turn.

   (a) **Net**: some acts will lead to both pleasure and pain. The Utilitarian thinks that the pain must be subtracted from the pleasure.

   (b) **Aggregate**: some acts will bring pleasure/pain to multiple people. The Utilitarian thinks that we must *aggregate* everybody's individual pleasures and pains to get one measure of the total amount of (net) pleasure. This is the good which we are attempting to maximize.

   (c) **Expected**: If we don't know for sure which state-of-affairs will result from an action, then we must take a probability-weighted average of each of them. This probability-weighted average is called an *expectation*. The (expected value) utilitarian thinks that it is right to choose the action which maximizes this expectation.

   (d) **Pleasure**: By 'pleasure', the utilitarian means something like 'happiness'—any enjoyable mental state counts as a pleasure. Not all pleasures are created equal—they differ in intensity and phenomenal character, at least. The Utilitarian assumes, however, that pleasure comes in degrees, and that all pleasures can be measured with a common scale.

      i. Let's call this common scale a *utility function*, $\mathcal{U}_i$ (this is the utility function for individual i). If we hand $\mathcal{U}_i$ a state-of-affairs, S, then $\mathcal{U}_i$ hands us back some number, $\mathcal{U}_i(S)$, which tells us how much pleasure individual i has in state-of-affairs S.

      ii. Terminology: in state-of-affairs S, individual i has $\mathcal{U}_i(S)$ *utiles*. (A *utile* is a unit of measurement, just like a *degree* in temperature, a *meter* in length, or a *second* in time.)

      iii. Note that, by assuming that there is such a common scale of measurement for pleasure, we've assumed that all pleasurable experiences are *comparable*. We can say whether, and to what degree, understanding a proof is more pleasurable than eating an ice cream cone.

      iv. Then, the utilitarian thinks that the overall goodness, $\mathcal{U}_G$, of a state-of-affairs, S, is given by just *summing up* the individual utilities of the people in that state-of-affairs.

$$\begin{aligned}\mathcal{U}_G(S) &= \mathcal{U}_1(S) + \mathcal{U}_2(S) + \mathcal{U}_3(S) + \cdots + \mathcal{U}_N(S) \\ &= \sum_{i=1}^{N} \mathcal{U}_i(S)\end{aligned}$$

8. Consider the following three cases.

   (a) At the local hospital, there are five very sick patients who need organ transplants in order to survive. Tom—who has no relatives or job, and who the doctors know will not be missed—has come in to have his tonsils removed. When Tom is under anesthetic, the doctors painlessly kill him and remove his organs, distributing them to the five sick patients. The patients go on to lead lives which are each just as happy as the life Tom would have led, had the doctors only removed his tonsils.

   (b) Every day, Bill the bully beats up Vince the victim. When Sam learns of this, he intervenes, standing up to the bully and telling him to leave Vince alone. Bill (predictably) beats up Sam instead. Vince was quite used to being beat up by Bill, while Sam is new to the experience, so Sam is made much less happy by the beating than Vince would have been.

   (c) The outcome of the national election will be the same whether Daniel votes or not. Waiting in line makes Daniel unhappy. He'd be happier staying at home. So Daniel stays at home and doesn't vote. He lies to everyone about this, so that nobody else knows that Daniel didn't vote.

9. In the first case, utilitarianism says that the doctors acted rightly. In the second case, utilitarianism says that, by standing up to the Bully, Sam acted wrongly. In the third, utilitarianism says that, by staying home and not voting, Daniel acted rightly. Insofar as we find these consequences objectionable, this gives us reason to worry about utilitarianism as a theory of right action.

2.2 Kant's Moral Theory

> *What does Kant's moral theory say? Illustrate Kant's moral theory by explaining how it could be used to show that cheating on an exam is wrong.*

10. Consider again the case of Daniel: what kinds of things might we say to explain why Daniel acted wrongly (if we think he did)? A common refrain is the following: "What if everyone stayed home instead of voting?"

    (a) Note: Daniel can agree that it would be bad if *everyone* stayed home instead of voting. But he knows that that won't happen. And, since no one knows that he didn't vote, he knows that his not voting doesn't make it any more or less likely that others won't vote.

    (b) When we point to the possibility of *everyone* not voting, we're not saying that this is likely to come about, nor that Daniel's action might play some role in bringing it about. But we still feel that this possibility can tell us something about how what Daniel has done is wrong.

11. Here's one way of developing this thought:

    Rule Utilitarianism The goodness of a system of rules is given by the amount of net, aggregate happiness that would result from everyone trying to follow those rules. An act is right iff it conforms to the *best* system of rules.

    (a) This is a kind of consequentialism. For, according to this theory, the right is determined by the good.

    (b) If we had a system of rules which permitted doctors to harvest the organs of their patients without their consent, nobody would go to the doctors. This would be worse than a state in which doctors required the consent of their patients in order to take organs. So the best system of rules will say that it's wrong to kill Tom and harvest his organs. (Think about what the theory would say about the other cases).

12. A worry about rule utilitarianism: it collapses back down to regular (act) utilitarianism.

    (a) Consider any system of rules which *always* forbids harvesting organs without the consent of the patients. Take that system of rules and emend it so that it includes the following opt-out clause: if Tom comes in to get his tonsils removed on September 17th, 2018, and there are five sick patients in need of organs, and you are certain that no one will find out, then remove Tom's organs and distribute them to the five.

    (b) If the doctors followed *this* system of rules, thing would be better off. And we can build in similar opt-out clauses for any act which promotes happiness in any particular case. So why doesn't rule utilitarianism just end up 'collapsing' back down to regular (act) utilitarianism?

13. Here's another way of developing the same thought (due to Immanuel Kant):

    Kant's Moral Theory Your act is morally right iff you can consistently will that the maxim on which you act can be universally followed.

    (a) A *maxim* is a general rule on the basis of which you act. Kant thinks that, whenever you act, you have some implicit maxim guiding the action.

    (b) If you can consistently will that everyone acts in accord with your maxim, then your act is morally permissible. If you cannot consistently will that everyone acts in accord with your maxim, then your act is not morally permissible.

    (c) Note an important difference between Kant's moral theory and rule utilitarianism. Both Kant and the rule utilitarian consider a possibility in which everyone follows a rule (or *maxim*). However, when considering that possibility, the rule utilitarian asks: 'how good is it?'. Kant, in contrast, asks: 'could you consistently will that possibility to be actual?'. On Kant's view, the rightness of the action isn't determined by the *goodness* of this possibility. So his theory is non-consequentialist.

14. An example:

    (a) You need money, but you know that you will be unable to repay a loan. Even so, you ask for money, promising to repay it. You do so on the basis of the maxim 'if I need money, then I will make a promise to repay a loan, even if I won't be able to'.

    (b) First, we 'generalize' your maxim, so that it applies not only to you, but to everyone else as well: 'if anyone needs money, then they will make a promise to repay a loan, even if they won't be able to'.

    (c) If this maxim were universally followed, then lenders would stop trusting that their loans will be repaid, and they will stop lending.

    (d) You cannot *consistently will* for this situation to be actual, for two reasons: 1) if there were no lenders, then nobody would be able to follow the maxim. So the situation in which everyone follows the maxim is contradictory. (This is a *contradiction in conception*). Also note that: 2) when you act, you will to obtain money; but, if your maxim were universally followed, you would not obtain money. For this reason, also, you cannot will that your maxim will be universally followed. (This is a *contradiction in will*.)

Chapter 3

Rationality

3.1 Expected Utility Theory

> *Say what a "folk psychological" explanation of actions consists in (which factors are cited to explain the action?). Then, introduce the technical surrogates for these factors which appear in expected utility theory. Precisely characterize which actions are rational according to expected utility theory.*

1. So-called 'folk psychology' explains people's behavior by appealing to their *beliefs* and their *desires*.
   (a) Why did Bob invest in Apple? Because he *believes* that Apple's stock will increase in value, and he *desires* to make money by selling the stock at a higher value than he bought it.
2. This explanation supposes that Bob is *rational*—that is, that, given his beliefs and desires, his actions *make sense*. We presuppose that Bob is going to act in the way that is most likely to satisfy his desires, given his beliefs.
3. A full explanation must also include Bob's *level of confidence* that Apple stock will increase, and note the fact that Bob desires collecting a return on his investment later on *more* than he desires spending the money on himself now.
4. So, we can introduce a more complicated technical surrogate for belief and desire:
   (a) *subjective probability* (or *degree of belief*, or *credence*); and
   (b) *subjective utility* (or *degree of desire*)
5. Let's decide that we're going to represent an individual's *subjective probability* with a function, $\mathcal{P}$.
   (a) We hand $\mathcal{P}$ a state-of-affairs, S, and it hands us back a real number between 0 and 1. This number, $\mathcal{P}(S)$, represents how *likely* the individual thinks it is that that state-of-affairs obtains.
      i. If $\mathcal{P}(S) = 1$, then they are certain that S obtains.
      ii. If $\mathcal{P}(S) = 0$, then they are certain that S does not obtain.
      iii. If $\mathcal{P}(S) = 0.5$, then they think that S is as likely to obtain as not.

   We'll assume that $\mathcal{P}$ is a probability function.

6. And let's represent an individual's *subjective utilities* with a function, $\mathcal{U}$. We hand $\mathcal{U}$ a state-of-affairs, S, and it hands us back some real number—any real number.

   (a) This number, $\mathcal{U}(S)$, represents the degree to which the individual *desires* that the state of affairs S obtains.

   (b) If $\mathcal{U}(S) > \mathcal{U}(T)$, then they desire that S obtain more than they desire that T obtain.

      i. Note that there is a *very big difference* between the way that we are using 'utility' today and the way that we used 'utility' last week when discussing utilitarianism.

      ii. For the utilitarian, the *utility* of a state was the net amount of *pleasure* someone experienced in that state.

      iii. We are now assuming that the *utility* of a state is a measure of the degree to which someone *desires* that state.

      iv. These two can come apart. [Try to think of a scenario in which someone desires something which causes them pain, or in which someone experiences pleasure from something they don't desire.]

7. Expected Utility Theory supposes that we may think of *rational* individuals as having a certain *probability* function and a certain *utility* function; and, it supposes further that an action is rational to perform iff it *maximizes* expected utility.

8. For instance, suppose that, for you, the utility of $\$x$ is x, for any x. And suppose that I'm about to flip a fair coin, and I offer you the following wager, which you are free to either accept or reject:

\$10	if the coin lands heads
$-\$9$	if the coin lands tails

   (a) We may represent your *decision problem* here by separating out three different factors: i) the possible *outcomes*; ii) the possible *actions*; and iii) the *states of the world* which determine which outcome results from which action.

      i. In this case, the possible *outcomes* are:

         A. you get \$0 (if you don't take the wager);

         B. you get \$10 (if you take the wager and win); and

         C. you lose \$9 (if you take the wager and lose).

      ii. The possible *actions* are:

         A. you take the wager; and

         B. you don't take the wager

      iii. And the *states of the world* which determine which outcome a given action gets mapped to are:

         A. the coin lands heads; and

         B. the coin lands tails.

   (b) In general, we may display these components of your decision problem in a matrix like the

following:

$$\begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{ccccc} S_1 & S_2 & S_3 & \cdots & S_n \end{array} \\ \begin{bmatrix} O_{1,1} & O_{1,2} & O_{1,3} & \cdots & O_{1,n} \\ O_{2,1} & O_{2,2} & O_{2,3} & \cdots & O_{2,n} \\ O_{3,1} & O_{3,2} & O_{3,3} & \cdots & O_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O_{m,1} & O_{m,2} & O_{m,3} & \cdots & O_{m,n} \end{bmatrix} \end{array}$$

(c) In this particular problem, we get a matrix like this:

$$\begin{array}{c} \\ \text{take the wager} \\ \text{refuse the wager} \end{array} \begin{array}{c} \begin{array}{cc} \text{heads} & \text{tails} \end{array} \\ \begin{bmatrix} \$10 & -\$9 \\ \$0 & \$0 \end{bmatrix} \end{array}$$

(d) We may then calculate the *expected utility* of an action A_i, '$\text{Exp}[\mathcal{U}(A_i)]$', by going along the row associated with that action, and, for each state of affairs S_j, *weighting* the utility of the outcome $O_{i,j}$ by the *probability* that state S_j is actual.[1]

$$\text{Exp}[\mathcal{U}(A_i)] \stackrel{\text{def}}{=} \sum_{j=1}^{n} \mathcal{U}(O_{i,j}) \cdot \mathcal{P}(S_j)$$

Alternatively, assuming $O_{i,j} = A_i \& S_j$, we could write that

$$\text{Exp}[\mathcal{U}(A_i)] \stackrel{\text{def}}{=} \sum_{j=1}^{n} \mathcal{U}(A_i \& S_j) \cdot \mathcal{P}(S_j)$$

Or, if we don't want to worry about all messy subscripts, we could just write:

$$\text{Exp}[\mathcal{U}(A)] \stackrel{\text{def}}{=} \sum_{S} \mathcal{U}(A \& S) \cdot \mathcal{P}(S)$$

(Read this as follows: "the expected utility of A is the sum, over S, of the utility of A in the state S, multiplied by the probability of S.")

(e) In this particular case, the *expected utility* of taking the wager is:

$$\begin{aligned} \text{Exp}[\mathcal{U}(\text{take the wager})] &= \mathcal{U}(\$10) \cdot \mathcal{P}(\text{heads}) + \mathcal{U}(-\$9) \cdot \mathcal{P}(\text{tails}) \\ &= 10 \cdot 0.5 - 9 \cdot 0.5 \\ &= 5 - 4.5 \\ &= 0.5 \end{aligned}$$

and the *expected utility* of refusing the wager is

$$\begin{aligned} \text{Exp}[\mathcal{U}(\text{refuse the wager})] &= \mathcal{U}(\$0) \cdot \mathcal{P}(\text{heads}) + \mathcal{U}(\$0) \cdot \mathcal{P}(\text{tails}) \\ &= 0 \cdot 0.5 + 0 \cdot 0.5 \\ &= 0 \end{aligned}$$

[1] If we want to be especially careful, we should say that 'A_i' is actually the state-of-affairs of your performing the action, rather than the action itself. (That's because $\mathcal{P}$ and $\mathcal{U}$ are only defined for states-of-affairs, not actions.) But let's not worry about this complication.

(f) Expected Utility Theory, EUT, then makes two claims.

i. Firstly, *it is rationally required* for the degree to which you desire the act of taking the wager (that is, your utility for taking the wager) to be equal to the expected utility of taking the wager. Additionally, it is rationally required for your utility for refusing the wager to be equal to the expected utility of refusing the wager. That is:

$$\begin{aligned}\text{It is a requirement of rationality that: } & \mathcal{U}(\text{take the wager}) = \text{Exp}[\mathcal{U}(\text{take the wager})],\\ \text{and it is a requirement of rationality that: } & \mathcal{U}(\text{refuse the wager}) = \text{Exp}[\mathcal{U}(\text{refuse the wager})].\end{aligned}$$

ii. Secondly, *it is irrational* for you to perform an act when there is some other act which you desire more than it.

(g) Since, in this case,

$$\text{Exp}[\mathcal{U}(\text{take the wager})] > \text{Exp}[\mathcal{U}(\text{refuse the wager})]$$

The first claim of expected utility theory tells us that it is rational for you to desire taking the wager more than you desire refusing the wager,

$$\mathcal{U}(\text{take the wager}) > \mathcal{U}(\text{refuse the wager})$$

and then the second claim of expected utility theory says it is *rational* to take the wager, and it is *irrational* to refuse the wager.

9. More generally, EUT makes the following two claims.

Expected Utility Theory
Rationality requires that:
(a) for each act A, $\mathcal{U}(A) = \text{Exp}[\mathcal{U}(A)]$; and
(b) you do not perform an act if there is some other act with a higher utility than it.

3.2 Ordinal versus Cardinal Utility

Explain the difference between an ordinal utility function and a cardinal utility function. In particular, be sure to say when two ordinal utility functions are equivalent, and when two cardinal utility functions are equivalent. If it is possible, provide an example of a pair of utility functions which are ordinally equivalent but not cardinally equivalent. If this is not possible, then explain why it is not possible. If it is possible, provide an example of a pair of utility functions which are cardinally equivalent but not ordinally equivalent. If this is not possible, then explain why it is not possible.

10. Whenever we represent a quantity with a real number, you should want to know which features of the numbers are *meaningful*—or, to put the point another way, which features of those numbers *depends* upon the particular arbitrary representational system we've chosen, and which are *independent* of that representational system.

(a) Consider *zip code*. Zip codes assign numbers to regions of the country. However, the assignment is completely arbitrary. *None* of the features of those numbers have any meaning. We could just as well have given arbitrary English names to those regions. Just because my zip code is *greater* than yours, this doesn't means that the region of the country I live in is greater than yours, or (really) bears any interesting relation to the region of the country I live in.

(b) In contrast, consider the *place* you come in in a race. One runner gets 1 (if they came in 1st), another gets 2 (if they came in 2nd), and so on. These numbers tell us *something* about how fast the runners were, but it's not the case that the 2nd place runner was twice as slow as the 1st place runner. Nor that the difference in speed between the 2nd and the 3rd place runners is the same as the difference in speed between the 4th and the 5th place runners.

   i. *Place* is what's known as an *ordinal* scale. The only features of these numbers which are important is their order.

   ii. We could decide to say that the first person past the finish line came in '0^{th}-place', and the second person past the finish line came in '1^{st}-place', and so on, and we'd be saying exactly the same thing.

   iii. In general, two ordinal scales, $\mathcal{O}_1$ and $\mathcal{O}_2$, are equivalent (in the sense that they say exactly the same things) iff, for any inputs A, B,

$$\mathcal{O}_1(A) \geqslant \mathcal{O}_1(B) \quad \text{iff} \quad \mathcal{O}_2(A) \geqslant \mathcal{O}_2(B)$$

(c) Finally, consider the numbers we assign to temperatures. We may say that the temperature is 1° *Celsius*, or that it is 33.8° *Fahrenheit* (these are the same temperature). So the particular number we happen to use is arbitrary; however, temperatures are not like zip codes. Not all features of the numbers are arbitrary. In particular, the relations of *greater than* and *less than* make sense for temperatures. However, it does not make sense to say that one temperature is *twice as hot* as another. These properties of the numbers are not meaningful. For while 2° Celsius is twice 1° Celsius, 35.6° Fahrenheit (which is the same as 2° Celsius) is not twice 33.8° Fahrenheit (which is the same as 1° Celsius).

Of course, temperature scales encode more information than just the *order* of the temperatures. But the *ratios* of temperatures are not meaningful (*e.g.*, the fact that 2° Celsius is twice 1° Celsius is not meaningful). So what *is* meaningful?

   i. The features of temperature scales which are meaningful are 1) their order; and 2) the *ratios of differences*. That is, if $\mathcal{T}_1$ is one temperature scale, and $\mathcal{T}_2$ is another, then, for any objects with temperatures A, B, C, D, 1) $\mathcal{T}_1(A) \geqslant \mathcal{T}_1(B)$ iff $\mathcal{T}_2(A) \geqslant \mathcal{T}_2(B)$ (that is: $\mathcal{T}_1$ and $\mathcal{T}_2$ are *ordinally equivalent*); and 2)

$$\frac{\mathcal{T}_1(A) - \mathcal{T}_1(B)}{\mathcal{T}_1(C) - \mathcal{T}_1(D)} = \frac{\mathcal{T}_2(A) - \mathcal{T}_2(B)}{\mathcal{T}_2(C) - \mathcal{T}_2(D)}$$

   ii. That's because temperature is what's known as an *interval scale*. For interval scales, the zero and the one values are arbitrary, but all other features of the numerical representation are not. That means that, if you have one temperature scale $\mathcal{T}$, then you may change the position of the 1 value by multiplying $\mathcal{T}$ by any positive number you like, and you may change the position of the 0 value by adding any number (positive or negative) you like, and you'll get another scale which is *equivalent* to the first, in the sense that it says all the same things that the original scale said.

iii. So, if $\mathcal{T}_1$ is a correct measure of temperature, then, for any $\alpha > 0$ and any β,

$$\mathcal{T}_2(\cdot) = \alpha\,\mathcal{T}_1(\cdot) + \beta$$

is *also* a correct measure of temperature. (I'm writing '$\mathcal{T}_1(\cdot)$' to stand for the function $\mathcal{T}_1$. The '$\cdot$' is the blank space where you can put any of $\mathcal{T}_1$'s inputs.)

iv. If that's so, then say that $\mathcal{T}_1$ is *unique up to positive linear transformation* (or, sometimes, people say *unique up to positive affine transformation*).

11. Bearing all of this in mind, let's distinguish two kinds of utility functions:

ORDINAL UTILITY FUNCTION

An *ordinal utility function* measures desire on an ordinal scale. That is, it only encodes information about the *order* in which you desire states-of-affairs. Therefore, any two utility functions $\mathcal{U}_1, \mathcal{U}_2$ which are such that, for all states-of-affairs S, T,

$$\mathcal{U}_1(S) \geqslant \mathcal{U}_1(T) \quad \text{iff} \quad \mathcal{U}_2(S) \geqslant \mathcal{U}_2(T)$$

are equivalent ordinal utility functions.

CARDINAL UTILITY FUNCTION

A *cardinal utility function* measures desire on an interval scale. That is, it encodes information about the*order* and the *ratios of differences* of the degrees to which you desire various outcomes. The placement of 1 and 0 in a cardinal utility function are meaningless, but all other features of the numerical representation *are* meaningful. Or, to put it another way, iff you can find some positive α and some β such that, for all states-of-affairs S,

$$\mathcal{U}_2(S) = \alpha \cdot \mathcal{U}_1(S) + \beta$$

then $\mathcal{U}_1$ and $\mathcal{U}_2$ are equivalent cardinal utility functions.

3.3 Interpreting Utilities

12. According to Expected Utility Theory, agents have *subjective utilities* (degreed desires) and *subjective probabilities* (degreed beliefs), and they are *rational* iff they choose actions which *maximize expected utility*.

13. But what do we mean when we say that agents have utilities? In virtue of what do agents have the subjective utilities that they do?

    (a) *The representationalist* says that, fundamentally, talk of subjective utilities is elliptical for talk about *preference*.

    (b) To say that $\mathcal{U}_i$ is individual i's utility function is just to say that i has preferences which may be *represented* with $\mathcal{U}_i$.

14. Instead of starting with subjective probabilities and subjective utilities, let's instead start with a single binary relation: *preference*.

    (a) For any given agent—let's take you, for instance—we assume that that agent prefers some states of affairs to others.

(b) Write '$S \succ T$' to mean that you *strictly prefer* S to T.

$$S \succ T \quad \overset{\text{means}}{=} \quad \text{you strictly prefer } S \text{ to } T$$

To say that you *strictly* prefer S to T is to say at least that, given a choice between S and T, you are disposed to choose S, and not at all disposed to choose T. Alternatively: there's some amount you are disposed to pay to exchange T for S. That is: we give you T and ask you how much you're willing to pay to have S instead. If you strictly prefer S to T, then you'll name some positive amount.

(c) In contrast, you might be *indifferent* between S and T. We'll notate this with '$S \sim T$'.

$$S \sim T \quad \overset{\text{means}}{=} \quad \text{you are indifferent between } S \text{ and } T$$

How should we characterize indifference? Perhaps like this: you are disposed to pay nothing to exchange T for S and you are disposed to pay nothing to exchange T for S. That is: we give you T, and ask you how much you're willing to pay to have S instead. If you're indifferent between S and T, then you'll say: 'nothing'. Likewise, we give you S and ask you how much you're willing to pay to have T instead. If you're indifferent between S and T, then you'll say 'nothing'.

(d) Given the relations of strict preference and indifference, we may define a third, which we can call *weak preference*. We'll notate it with '$\succeq$'.

$$S \succeq T \quad \overset{\text{means}}{=} \quad \text{you weakly prefer } S \text{ to } T$$

We could then *define* weak preference in terms of strict preference and indifference, as follows:

$$S \succeq T \quad \overset{\text{def}}{=} \quad \text{Either } S \succ T \text{ or } S \sim T$$

To say that you *weakly* prefer S to T is to say at least that you are disposed to pay nothing to exchange T for S. That is: we give you S and ask you how much you're willing to pay to have T instead. If you weakly prefer S to T, then you'll say 'nothing'.

(e) Actually, we don't have to define things this way around. We could instead take *weak* preference as basic, and define both strict preference and indifference in terms of it. (See the problem set.)

15. Suppose that you have a weak preference relation $\succeq$ over states of affairs which satisfies the following three constraints:

TOTALITY
for any states of affairs S, T,

$$\text{Either } S \succeq T \text{ or } T \succeq S$$

REFLEXIVITY
For any state of affairs S,

$$S \succeq S$$

TRANSITIVITY
For any states of affairs S, T, U,

$$\text{if } S \succeq T \text{ and } T \succeq U, \text{ then } S \succeq U$$

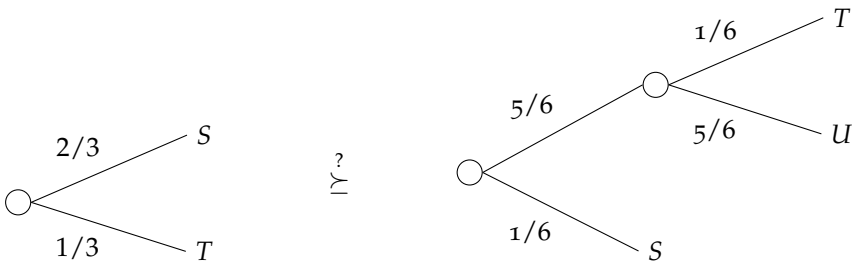
Then, according to the ORDINAL REPRESENTATION THEOREM, we may represent your preferences with an *ordinal* utility function.

ORDINAL UTILITY REPRESENTATION THEOREM
If you have a total, reflexive, and transitive preference ordering $\succeq$ over states of affairs, then you may be represented with an *ordinal* utility function $\mathcal{U}$ such that, for any states of affairs S, T,

$$\begin{aligned} & \mathcal{U}(S) \geqslant \mathcal{U}(T) \quad \text{iff} \quad S \succeq T \\ & \mathcal{U}(S) > \mathcal{U}(T) \quad \text{iff} \quad S \succ T \\ \text{and} \quad & \mathcal{U}(S) = \mathcal{U}(T) \quad \text{iff} \quad S \sim T \end{aligned}$$

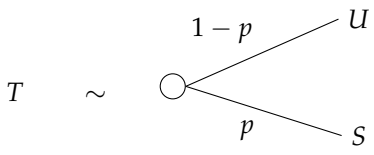
16. And suppose that you have a preference relation $\succeq$ defined over states of affairs *and arbitrary lotteries involving those states of affairs, and lotteries involving lotteries involving states of affairs, and so on and so forth*.

That is, you are no longer only opinionated about which *states of affairs*—S or T—are better than each other, but additionally you're opinionated about which *lotteries* are preferable to which other *lotteries*. For instance, you must have an opinion about whether you prefer a lottery that delivers S with probability $2/3$ and T with probability $1/3$ to a lottery which delivers S with probability $1/6$ and, with probability $5/6$, *another* lottery which delivers T with probability $1/4$ and U with probability $3/4$.



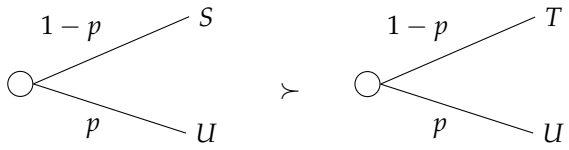
Suppose further that your preferences satisfy TOTALITY, REFLEXIVITY, and TRANSIVITY; and, in addition, they satisfy the following constraints:

CONTINUITY
For any S, T, U, if $S \succeq T$ and $T \succeq U$, then there is some probability p such that you are indifferent between T and the lottery which gives S with probability p and U with probability $1 - p$.



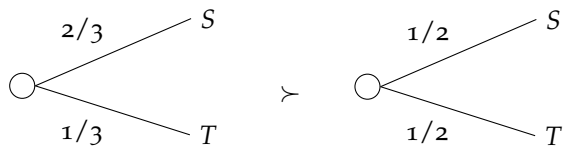
SWEETENING
If all else is held fixed, then you prefer replacing one potential outcome in a lottery with a

preferred outcome. (Sweetening the outcome on one branch makes a lottery preferable, all else held fixed). For instance, if $S \succ T$, then



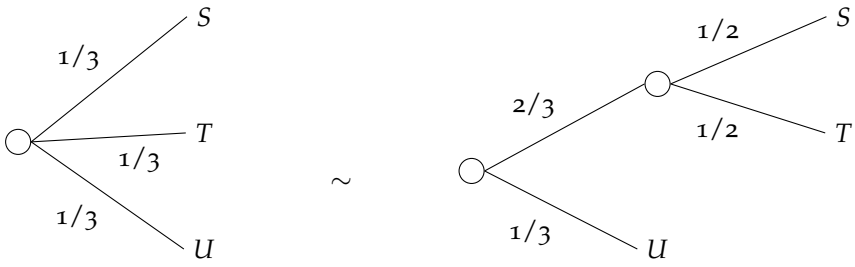
BETTER CHANCES

All else equal, a lottery is preferred if it has a higher probability of delivering the most preferred outcome. For instance, if $S \succ T$, then



REDUCTION OF COMPOUND LOTTERIES

You are indifferent between a lottery which has other lotteries as potential outcomes and a single lottery with the same final outcomes and the same final probabilities. For instance, you are indifferent between the following two lotteries:



If you satisfy all of these axioms, then the CARDINAL UTILITY REPRESENTATION THEOREM assures us that we may represent your preferences with a *cardinal* utility function—and, what's more, that you will be representable as an *expected utility maximizer*.[2]

CARDINAL UTILITY REPRESENTATION THEOREM

If you have a total, reflexive, and transitive preference ordering $\succeq$ over states of affairs and lotteries of states of affairs—and, in addition, $\succeq$ satisfies CONTINUITY, SWEETENING, BETTER CHANCES, and REDUCTION OF COMPOUND LOTTERIES, then you may be represented with a *cardinal* utility function $\mathcal{U}$, and a probability function $\mathcal{P}$ such that, for any two states of affairs S, T,

$$\begin{aligned} & \mathcal{U}(S) \geqslant \mathcal{U}(T) \quad \text{iff} \quad S \succeq T \\ & \mathcal{U}(S) > \mathcal{U}(T) \quad \text{iff} \quad S \succ T \\ & \mathcal{U}(S) = \mathcal{U}(T) \quad \text{iff} \quad S \sim T \\ \text{and} \quad & U(S) = \text{Exp}[\mathcal{U}(S)] \end{aligned}$$

[2]For the details and proofs of all of these claims, see the optional reading from Resnik.

That is: if you satisfy these constraints, they you may be represented as an *expected utility maximizer*. (Note: because $\mathcal{U}$ is a *cardinal* utility function, it is only unique up to positive linear transformations.)

Chapter 4

Rationality, day 2

4.1 Evaluating Expected Utility Theory

1. Distinguish two uses of expected utility theory:
   (a) a *descriptive* use, in which it merely says something about what people *will* do
   (b) a *normative* use, in which it gives advice about what you (in some sense) *ought* to do (what it would be *rational* to do
2. Notice that we could object to the descriptive use of expected utility theory without objecting to its normative use. So too could we object to its normative use without objecting to its descriptive use. The objections we will be discussing today may be understood as either objections to the normative use of EUT or to the descriptive use of EUT, though I'll focus on the objections to its *normative* use throughout.
3. Notice also that, on its own, expected utility theory does *not* say that:
   (a) It is irrational to prefer less money (for you) to more money (for you)
   (b) It is irrational to prefer that others get benefits at your expense
   (c) It is irrational to care about your *duties* or others *rights*

4.1.1 The Allais Problem

1. Suppose that I'm going to roll a 100-sided die, and you have a choice between the following two lotteries:

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2-10 & 11-100 \end{array} \\ \begin{bmatrix} \$1,000,000 & \$1,000,000 & \$1,000,000 \\ \$0 & \$5,000,000 & \$1,000,000 \end{bmatrix} \end{array}$$

   Which do you prefer?

2. Suppose that I'm again going to roll a 100-sided die, and you have a choice between the following two lotteries:

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2-10 & 11-100 \end{array} \\ \begin{bmatrix} \$1,000,000 & \$1,000,000 & \$0 \\ \$0 & \$5,000,000 & \$0 \end{bmatrix} \end{array}$$

Which do you prefer?

   (a) Many people prefer A to B, and prefer D to C. If you have these preferences (and you care only about money), however, then you are irrational according to EUT.

      i. Why?

      ii. Does this assume that your utilities are linear in dollars?

      iii. What does it assume about your utilities?

3. We may use this case to mount the following argument against EUT:

   P1) EUT says that the Allais preferences are irrational.

   P2) The Allais preferences are not irrational.

   C) EUT is false.

4. The problem, according to objectors, is that it seems to be rational to care about the fact that, with A, you are *guaranteed* to get $1,000,000. But the security of a sure $1,000,000 is not the kind of thing that expected utility theory allows you to take into consideration when making your decision (assuming, that is, that you value only money).

   (a) A reply: perhaps what you *really* care about is not *only* money. Perhaps you *also* care about avoiding the *regret* of having passed up a guaranteed $1,000,000. Or perhaps you *also* care about the *feeling* of security that comes with not having to worry about the outcome of the die roll.

   (b) How does this help?

4.1.2 The Ellsberg Problem

1. Suppose that I have an urn which contains 30 red balls, and 60 other balls, which are either yellow or green. You don't know anything about the distribution of green and yellow balls, other than that the total number of green and yellow balls is 60. There could be 60 green and 0 yellow, or 50 yellow and 10 green, or 30 green and 30 yellow. I am going to draw a ball at random from the urn, and I offer you the following two gambles, whose payouts depend upon the color of the ball selected:

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{ccc} \text{red} & \text{yellow} & \text{green} \end{array} \\ \begin{bmatrix} \$100 & \$0 & \$0 \\ \$0 & \$100 & \$0 \end{bmatrix} \end{array}$$

Which do you prefer?

Suppose next that everything is the same as before, except I offer you the following two gambles:

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} \text{red} & \text{yellow} & \text{green} \end{array} \\ \begin{bmatrix} \$100 & \$0 & \$100 \\ \$0 & \$100 & \$100 \end{bmatrix} \end{array}$$

Which do you prefer?

(a) Many people prefer A to B, but also prefer D to C. But this, too, is in violation of EUT.

i. Why?

ii. Does this assume that your utilities are linear in dollars?

iii. What does it assume about your utilities?

2. We may use this case to mount the following argument against EUT:

P1) EUT says that the Ellsberg preferences are irrational.

P3) The Ellsberg preferences are not irrational.

C) EUT is false.

3. The problem according to objectors, is that it seems rational to have different attitudes towards *risk* and *uncertainty*—where *risk* is a situation in which there are some known objective probabilities, and *uncertainty* is a situation in which there are no known objective probabilities.

4. Because expected utility theory always supposes that you have a single unique probability function, it does not recognize differences between risk and uncertainty.

4.1.3 Act-State Dependence

1. Here is a general decision-theoretic principle which follows from expected utility theory:

DOMINANCE
If the outcomes associated with the action A are preferred to the outcomes associated with the action B in *every* state of the world, then A should be preferred to B.

(a) An illustration:

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} S_1 & S_2 \end{array} \\ \begin{bmatrix} 50 & -100 \\ 40 & -105 \end{bmatrix} \end{array}$$

In every state of affairs, A's outcome is preferred to B's. So, the DOMINANCE principle says that A should be preferred to B.

(b) Why does this follow from EUT?

2. Suppose that 90% of people like you get cancer in their lifetimes. There is a pill which prevents the development of cancer with very high probability (99%). It costs $5. Should you buy it?

(a) Here's an argument that you shouldn't: Either you will get cancer or you won't. If you get cancer, then you would prefer to have your \$5. If you don't get cancer, then you would prefer to have your \$5. Suppose, then, that your (cardinal) utility function is like this:

$$\begin{array}{c} \\ \text{buy pill} \\ \text{don't buy pill} \end{array} \begin{array}{c} \begin{array}{cc} \text{cancer} & \text{no cancer} \end{array} \\ \begin{bmatrix} -1001 & 1000 \\ -1000 & 1001 \end{bmatrix} \end{array}$$

Then, not buying the pill *dominates* buying the pill. So the expected utility of not buying the pill is higher than the expected utility of buying it. So you shouldn't buy.

3. Problem: there is *act-state-dependence*. Which act we take *affects* which state of the world obtains. We shouldn't use expected utility theory when there is act-state dependence.

4. Richard Jeffrey's solution: we shouldn't maximize *expected utility*, but rather what Jeffrey calls *evidential expected utility*.

EVIDENTIAL EXPECTED UTILITY THEORY

Rationality requires that:

(a) for each act A, $U(A) = \sum_S P(S \mid A) \cdot U(S \,\&\, A)$; and

(b) you do not perform an act if there is some other act with a higher utility than it.

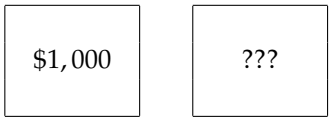
The Newcomb Problem

1. Suppose you find yourself in the following decision problem:

NEWCOMB PROBLEM

You awake in a room. Brain scientists tell you that you have been selected for an experiment. They have been developing technology which allows them to make quite accurate predictions about how people will behave in various decision scenarios. Three days ago, they abducted you and took a scan of your brain. On the basis of that brain scan, they made a prediction about what you will do today. In past trials, they have made the right prediction about 75% of the time.

In front of you are two boxes, one which is transparent, and one which is opaque. Within the transparent box, you can see \$1,000.



The scientists tell you the following (and you believe them): "you may either take both boxes, or only the mystery box on the right. The choice is up to you. But know this: we made a prediction about which box you would take. If we predicted that you would take *just* the mystery box, then we put \$1,000,000 in the mystery box. If we predicted that you would take *both* boxes, then we left the mystery box empty." Given that you take only the mystery box, the probability that it was predicted that you would take only the mystery box is 75%. Given that you take both boxes, the probability that it was predicted that you would take both boxes is 75%. Your utilities are linear in dollars (that is, the utility of receiving \$$x$ is x).

What do you do? Do you take just the mystery box, or do you take both boxes?

2. What does Jeffrey's *evidential* expected utility theory say you ought to do?

Chapter 5

Theories of Welfare and the Pareto Principle

What is a philosophical theory of welfare *(or* well-being*) a philosophical theory* of*? (That is: what is the goal of a philosophical theory of welfare? What theory of welfare is the Utilitarian committed to (and what does this theory say)? What is an objection which has been raised against this theory? What theory of welfare is the Welfare Economist committed to (and what does this theory say)? What is an objection which has been raised against this theory?*

5.1 Preview: the Ethical Theory of Welfare Economics

1. Recall that we definied Utilitarianism as the conjunction of three evaluative claims: firstly, a claim about *well-being*, or *welfare*; secondly, a claim about *goodness*; and thirdly, a claim about *rightness*.[1]

   Utilitarianism

   A *Utilitarian* is somebody who accepts each of the following claims.

   Hedonism: What it is for an individual to have greater welfare—*i.e.*, what it is for things to go well for an individual—is for them to have more pleasure and less pain.

   Welfarism: The goodness of a state of affairs is determined by the welfare of the individuals in that state-of-affairs.

   Consequentialism: The rightness of an action is determined by the goodness of the states-of-affairs which result from the action's performance.

2. The reason that it is helpful to separate out these three different commitments of Utilitarianism is that the ethical theory of welfare economics that we will be studying this semester agrees with Utilitarianism about *Welfarism* and *Consequentialism*, but disagrees with it about *Hedonism*. In place of Hedonism, the Welfare Economist accepts *Preferentism*—the theory according to which what it is for things to go well for you is for more of your *preferences* to be satisfied.

   Welfare Economics

   A *Welfare Economist* is somebody who accepts each of the following claims.

[1] A warning: this is not the way that Hausman, McPherson, and Satz define Utilitarianism. My definition outlines the position of classic Utilitarians like Bentham and Mill, which is important to cleanly separate from the position of modern welfare economists.

PREFERENTISM: What it is for an individual to have greater welfare—*i.e.*, what it is for things to go well for an individual—is for more of their preferences to be satisfied.

WELFARISM: The goodness of a state of affairs is determined by the welfare of the individuals in that state-of-affairs.

CONSEQUENTIALISM: The rightness of an action is determined by the goodness of the states-of-affairs which result from the action's performance.

3. To understand this difference, let's take a look at the different philosophical theories of *welfare*.

5.2 Welfare

5.2.1 The Goal of a Philosophical Theory of Welfare

1. Your level of *well-being*, or *welfare*, is determined by how well things are going for you. It is the thing you wish for those you love. It is the thing you wish to raise with reward, and the thing you wish to lower with punishment. It is the thing you give up with sacrifice, and the thing you hoard with selfishness.

   (a) This concept is individualistic. In asking about *Sabeen's* welfare, we are not asking about how well things are going for Matthew, or David, or anybody else. We are simply talking about how well things are going for Sabeen.

   (b) The concept is not moralistic. As a conceptual matter, it could be that things are going well for Sabeen, even though they shouldn't. As a conceptual matter, it could be that acting wrongly sometimes makes somebody better off.

2. The task of providing a philosophical theory of welfare is the task of say what welfare *consists in*. It is the task of saying *in virtue of what* a life is going better or worse.

   (a) In providing a philosophical theory of welfare, we don't wish to say what kinds of things *cause* a life to go better or worse. (Perhaps it's money; perhaps it's power; etc.) Rather, we wish to say what power or money would have to cause, were they to cause your life to go better.

   (b) That is, we wish to give an account of what is *intrinsically* good for a person; not what is *instrumentally* good for a person.

      i. Something is *intrinsically* good for a person if that thing is good for its own sake. (To decide whether something is intrinsically good, ask yourself: would it be a good thing to have, even if it didn't have any causal consequences whatsoever?)

      ii. Something is *instrumentally* good for a person if that thing is good in virtue of bringing about consequences which are good for that person.

      iii. Question: is it possible for something to be both intrinsically and instrumentally good?

5.2.2 Philosophical Theories of Welfare

3. The first kind of theory of welfare we will consider is an *objective list* theory.

OBJECTIVE LIST

There are a collection of objective goods—health, happiness, friendship, love, knowledge, aesthetic appreciation, *e.g.*—which are objectively good for people. A life is going well (at a time) just to the extent that it possesses the items on the list (at that time).

(a) A concern: how do we weigh the different items on the list? Might there be incomparabilities?

(b) An objection: what if I don't care about having friends, or knowledge, or aesthetic appreciation? The objective list theory says that things are going bad for me if I don't appreciate music, even though I don't at all *want* to appreciate music; even though music doesn't really do very much for me. I hate the opera and so give you the tickets. According to the objective list theory, I have made a sacrifice. But this can seem wrong—it doesn't seem like a sacrifice at all if opera leaves me cold.

(c) Another objection: the objective list theory says that you could *reward* me by providing me with a good on the list that I don't care for or about. For instance, it says that you could reward me by giving me health, friends, love, or the ability to appreciate music, even though I don't at all *want* to be healthy, to have friends, to love, or to appreciate music.

4. Another kind of theory of welfare is given by a *mental state* view—according to which your well-being is determined in some way or other by your mental state, and *only* your mental state. One particular mental state view is *hedonism*.

HEDONISM

A person's welfare (at a time) is determined by the amount of pleasure and the amount of pain in their life (at that time).

(a) An objection: Nozick's *experience machine*. Those in the experience machine experience greater pleasure than those outside it, but (unbeknownst to them), nothing in their life is real. Their loved ones are actually computer simulations. Their victories were guaranteed. You are supposed to feel that such a life is not really better for the person living it, even though they are receiving more pleasure than pain.

   i. Along the same lines: the *deceived businesswoman* believes that she is respected by her colleagues and that her husband and children love her. Unbeknownst to her, however, her colleagues think she is a joke and laugh about her in private. Her husband has been having an affair for years, and her children are only kind to her in order to get their inheritance. The *undecived businesswoman* is just like the deceived businesswoman, except that all of her beliefs are true—her colleagues really do respect her, her husband really does love her, and her children really do adore her.

   ii. Any mental state theory must say that the deceived and undeceived businesswoman have the same level of welfare. But it seems as though things are going better for the undeceived businesswoman than they are for the deceived businesswoman.

(b) Another objection: what about people who don't care about pleasure? Consider a rock climber who cares only for how many rocks she climbs, and not at all about how much pleasure or pain she receives. Hedonism says that things are going worse and worse for her as she climbs more and more rocks, because climbing rocks is painful. But you might think that things are going very well for this person—you might think that preventing her from climbing rocks would be no reward, and allowing her to climb rocks no punishment.

i. Along the same lines: consider a masochist who desires pain, and not pleasure. You might think that things are going well for the masochist when they get the pain they desire, and that giving them pain would be a way of rewarding them.

5. These objections lead naturally to a third kind of theory about welfare, which says that things are going well for you to the extent that your desires are satisfied.

   DESIRE SATISFACTION
   A person's welfare (at a time) is determined by how many of their desires are satisfied (at that time).

   (a) A clarification: when we say that your desires are *satisfied*, we don't mean that you yourself have any *sensation* of satisfaction. We mean merely that you want it to be the case that p (for whatever p you desire), and that p is in fact true. You needn't be *aware* that p is true in order for your desire that p to be satisfied. So desire satisfaction is not a mental state theory. Your welfare does not just depend upon your mental state. It additionally depends upon the way the world is.

   (b) A worry: we shouldn't just *count up* the number of desires that are satisfied; for desires come in degrees. Jessica wants a new bicycle, avacado toast, and for her partner to love her. But it is not better for Jessica to have the bike, the avacado toast, and her partner to loathe her than it is for her to lack the bike and avacado toast but have her partner's affection.

6. One way of solving this problem is by substituting out the coarse-grained, binary notion of *desire* for the more fine-grained, comparative notion of *preference*. For, as we have already seen, preferences can be used to make sense of the notion of *degreed* desire, or utilities. So one way of improving the desire satisfaction theory is by swapping out desires for preferences. (For reasons that will become clear below, we'll call this theory of welfare 'actual preferentism').

   ACTUAL PREFERENTISM
   A person's welfare (at a time) is determined by the degree to which their actual preferences are satisfied (at that time).

   (a) A clarification: again, when we speak of the degree to which your preferences are *satisfied*, we are not referring to any mental state of *satisfaction*. The theory of actual preferentism says that, if you prefer your children loving you to your children secretly despising you, then things are going better for you if they love you than they are if they secretly despise you. It does not matter whether you are *aware* that you children secretly despise you or not.

   (b) An objection: not all preferences are 'self-regarding'. Will give $100 to a charity that fights malaria in sub-Saharan Africa. This is good evidence that Will *prefers* the charity having those $100 to him having those $100 himself. So the actual preferentist theory says that Will's welfare is enhanced by his donating the $100. But this intuitively looks like a case of *sacrifice*.

7. A proposed fix:

   ACTUAL SELF-REGARDING PREFERENTISM
   A person's welfare (at a time) is determined by the degree to which their actual *self-regarding* preferences are satisfied (at that time).

(a) An objection: it doesn't seem that the Penguins fan's preference for the Pens winning the Stanley cup over the Predators winning the Stanley cup is self-regarding. However, it does seem that things were going well *for the Pens fan* when the Pens fan won the Stanley cup.

(b) Another objection (to both forms of actual preferentism): I prefer the fish to the chicken. However, unbeknownst to me, the fish is contaminated and will cause a weekend of vomiting. Actual preferentism seems to say that things go well for me when I eat the fish.

8. A proposed fix: don't consider your *actual* preferences, but rather consider your *informed* preferences—the preferences you *would* have, *were* you to be informed of all the relevant facts.

   INFORMED (SELF-REGARDING) PREFERENTISM
   A person's welfare (at a time) is determined by the degree to which their informed (self-regarding) preferences are satisfied (at that time).

   (a) Another objection: it looks as though the heroin addict *desires* to have heroin—that is, it looks as though their preferences would be satisfied to a large degree, were they to get heroin. But it does not seem that receiving the heroin would make them better off. It seems as though it would make them worse off.

9. A proposed fix: don't consider just your *informed* preferences, but also consider your (in some sense) *idealized* preferences. We consider what your preferences *would* be, not just if you were informed of all the relevant facts, but also if you were freed from your addictions, and if you were *idealized* in various other ways. It is *these* preferences which should be used to determine your level of well-being.

   IDEALIZED (SELF-REGARDING) PREFERENTISM
   A person's welfare (at a time) is determined by the degree to which their idealized (self-regarding) preferences are satisfied (at that time).

   (a) What does the process of idealization look like? We take you, as you are, and provide you with all true information, we equip you with all the concepts needed to properly appreciate the options you rank, we provide you with the experiences you need to experience in order to understand what it would be like to eat Vegemite, to lose a loved one, to climb Mount Everest, and so on and so forth.

   (b) We then elicit your (idealized, self-regarding) preferences.

   (c) Question: which *self* is at issue now? Are the preferences self-regarding in the sense that they regard *your now idealized self*. Or do they regard your *unidealized* self? Will this make a difference?

   (d) Question: as you are given new information, concepts, and experiences, there will be incremental changes in your preferences. What if there is *path-dependence*? What if the order in which you are idealized makes a difference to the preferences you end up having? Is there some *correct* way in which you should be idealized?

10. We will be talking much more about preferentism, its various versions, and various objections to it, next class.

11. Finally, we should note that this list of theories is not exhaustive. There are many other alternatives—perhaps welfare consists in the achievement of one's *aims*, perhaps it consists in *knowing* that your desires have been satisfied, perhaps it consists in the consumption of jelly—and there are (infinitely) more alternatives out there.

    (a) One alternative deserves note: we might endorse a kind of *hybrid* theory of welfare which mixes various theories we've already seen.

    (b) For instance, we might think that welfare consists in deriving *pleasure* from things like health, friendship, love, knowledge, aesthetic appreciation, *etc.*. This mixes the mental state theory of hedonism with the objective list theory, and does so in such a way that avoids some of the objections to both hedonism and the objective list theory. Can you think of any objections to this theory?

    (c) Alternatively: we might say that welfare consists having a satisfied desire for pleasures. This mixes a desire satisfaction theory with hedonism, and does so in such a way that it avoids some of the objections to both. Can you think of any objections to this theory?

12. For much of the remainder of this class, we will be focusing on preferentism, but it is important to keep in mind that, while this theory of welfare is assumed by the welfare economist, and while it may be the correct theory, it is but one theory among many. This assumption of welfare economics is substantive and contentious.

5.3 Preferentist Welfarism

First, say what the preferentist welfarist believes. Then, explain why the preferentist welfarist cannot evaluate states-of-affairs by simply summing the total amount of utility in those states of affairs. Next, explain how the Pareto Principle allows the preferentist welfarist to side-step this problem. Finally, state the first and second fundamental theorems of welfare economics.

1. The welfare economist accepts (some form of) *preferentism* about welfare. They additionally accept the thesis of *welfarism*,

   > Welfarism
   > The goodness of a state of affairs is determined by the welfare of the individuals in that state-of-affairs.

    (a) Let's pause to appreciate the following: in the previous two classes, we've been talking about the *prudential rationality* of preferences, and the prudential rationality of actions. That's an evaluative notion, but not a *moral* one. Though we are still talking about preferences, our focus has shifted. We are now evaluating the *goodness* of certain states-of-affairs. (And later still, once we add on the Consequentialism component of welfare economics, we'll be evaluating the *rightness* of various actions or policies.)

        i. Keep in mind: in saying something about the goodness of states of affairs, it is not *automatic* that the *acts* which achieve them are *right*. For, if a deontological theory is correct, then acting so as to bring about a good state of affairs may be *wrong*.

2. According to Preferentist Welfarism, we can determine goodness by looking at the degree to which individuals' preferences are satisfied. But that's rather vague. How *exactly* should we determine goodness?

5.3.1 Aggregate Utility?

3. Utilitarians had the idea of just *summing up* everyone's individual utility. Why not follow them on this, and evaluate the goodness of states of affairs in terms of the *aggregate* utility of those states of affairs?

   (a) Because there is a technical problem, which is redolent of the *utility monster* objection to utilitarianism.

   (b) Suppose that we have only two people, Ike and Janet. Ike's preferences are representable with the cardinal utility function $\mathcal{U}_i$, and Janet's preferences are representable with the cardinal utility function $\mathcal{U}_j$. And suppose that we wish to compare the states of affairs A, B, and C to say which is better. Ike and Janet's utilities in these states of affairs are shown below.

	$\mathcal{U}_i$	$\mathcal{U}_j$
A	100	10
B	60	60
C	10	100

   So, Ike strictly prefers A to B to C,

   $$A \succ_i B \succ_i C$$

   And Janet strictly prefers C to B to A

   $$C \succ_j B \succ_j A$$

   Now, if we attempt to determine the goodness of A, B, and C by *aggregating* the utilities of Ike and Janet, then we will just sum up the utilities of Ike and Janet in each of these states of affairs.

	$\mathcal{U}_i$	$\mathcal{U}_j$	$\mathcal{U}_G = \mathcal{U}_i + \mathcal{U}_j$
A	100	10	110
B	60	60	120
C	10	100	110

   And we would then say that B is better than both A and C, though A and C are equally good.

   $$B \succ_G A \sim_G C$$

   ('G' for *Group*, or for *Goodness*.)

   (c) But wait—these are only *cardinal* utility functions. So, if $\mathcal{U}_i$ represents Ike's preferences, then so too will

   $$\mathcal{U}_i' = 10 \cdot \mathcal{U}_i + 10$$

   But then, we would have

	$\mathcal{U}_i'$	$\mathcal{U}_j$
A	1010	10
B	610	60
C	110	100

And if we *then* just sum up the values in the rows, we'll get

	$\mathcal{U}_i'$	$\mathcal{U}_j$	$\mathcal{U}_G = \mathcal{U}_i' + \mathcal{U}_j$
A	1010	10	1020
B	610	60	670
C	110	100	210

And we'll end up saying that A is better than B , which is better than C—that is, our goodness ordering will match up perfectly with Ike's preference ordering.

$$A \succ_G B \succ_G C$$

We've essentially turned Ike into a utility monster, and his preference ordering has become the group preference ordering just by changing the units we use to measure Ike's preferences. We could do the same with Janet.

(d) This could be solved if we had some way of comparing the *units* of Ike and Janet's cardinal utility functions, and saying that they are the same; but how could that be accomplished?

i. Note that this may be turned into an argument *against* utilitarianism—for we now see how substantive the idea that we could just add up different people's utilities really is.

A. In order for the addition to make sense, and the ordering determined by that addition to be unique, we must be able to coordinate *both* the zero *and* the one of everyone's utilities scales. And this can look like a very strong assumption.

(e) This technical problem is known as the problem of making *interpersonal comparisons of utility*. The worry is that there is no way to compare the scale in which we represent Ike's utility and the scale in which we represent Janet's utility. And if there's no way to compare those scales, then it is unclear how we are supposed to determine goodness from preference in the way that the preferentist welfarist wants.

i. Can you think of a way to compare Ike and Janet's scales?

5.3.2 The Pareto Principle

4. The preferentist welfarist can take a different approach: try to start making evaluative claims about the goodness of states of affairs with just the following bare-bones notions (named after the Italian philosopher/economist Vilfredo Pareto):

Pareto Improvement
A state of affairs A is a Pareto Improvement on a state of affairs B iff *somebody* strictly prefers A to B *nobody* strictly prefers B to A.

$$(\forall i) A \succeq_i B \quad \& \quad (\exists i) A \succ_i B$$

Pareto Optimality
A state of affairs is Pareto Optimal iff there is no other possible state of affairs which is a Pareto Improvement on it.

Pareto Sub-optimality
A state of affairs is Pareto Suboptimal iff there is some other possible state of affairs which is a Pareto Improvement on it.

5. And we can endorse the following moral principle:

   Pareto Principle
   If A is a Pareto improvement on B, then A is strictly better than B.

   $$[(\forall i) A \succeq_i B \;\; \& \;\; (\exists i) A \succ_i B] \;\; \Rightarrow \;\; A \succ_G B$$

   (a) If we want to be careful, we should distinguish the Pareto Principle from a weakened and a strengthened version of the Pareto Principle:[2]

   Weak Pareto Principle
   If everybody strictly prefers A to B, then A is strictly better than B.

   $$(\forall i) A \succ_i B \;\; \Rightarrow \;\; A \succ_G B$$

   Pareto Principle
   If somebody strictly prefers A to B and nobody strictly prefers B to A, then A is strictly better than B.

   $$[(\forall i) A \succeq_i B \;\; \& \;\; (\exists i) A \succ_i B] \;\; \Rightarrow \;\; A \succ_G B$$

   Strong Pareto Principle
   If everybody is indifferent between A and B, then A is just as good as B,

   $$(\forall i) A \sim_i B \;\; \Rightarrow \;\; A \sim_G B$$

   and, if somebody strictly prefers A to B and nobody strictly prefers B to A, then A is strictly better than B.

   $$[(\forall i) A \succeq_i B \;\; \& \;\; (\exists i) A \succ_i B] \;\; \Rightarrow \;\; A \succ_G B$$

6. Note how minimal this principle is. *None* of the states of affairs A, B, or C in the previous section is a Pareto improvement on the others; so the Pareto Principle cannot allow us to choose between them.

   (a) In particular, notice that the betterness ordering over states-of-affairs we get from the Pareto Principle is not *total*. There will be many states of affairs, A, B for which the Pareto Principle will not tell us whether $A \succeq_G B$ or $B \succeq_G A$.

   (b) All that the Pareto Principle tells us is that *some* states-of-affairs are *bad*. They are bad in the sense that there is some other alternative state-of-affairs which everybody at least weakly prefers, and some strictly prefer. But the Pareto Principle does *not* tell us that Pareto optimal states of affairs are *best*, or even that they are particularly *good*.

   (c) It is consistent with the Pareto Principle that a Pareto optimal state-of-affairs is worse than some Pareto sub-optimal state of affairs.

   (d) What the Pareto Principle tells us is simply that the Pareto optimums *lack* a certain *bad-making* feature.

[2] Technical quibble: I'm assuming that every individual's preference ordering is *total*.

5.4 The Fundamental Theorems of Welfare Economics

1. The Pareto Principle, together with the so-called *fundamental theorems of welfare economics*, may be used to offer a defense of the free market. The first fundamental theorem goes as follows:

   FIRST FUNDAMENTAL THEOREM OF WELFARE ECONOMICS
   The state of affairs resulting from a *free market* will be Pareto Optimal.

2. In my statement of this theorem, I'm using the word 'free market' as a technical term. In order to have a free market, the following conditions must be met:

   (a) There is perfect information—everybody knows the price of all goods, and which goods are for sale, and so on and so forth.

   (b) There are no barriers to trade—People are free to trade any and all goods and services (including, *e.g.*, pollution, love, and organs), at no cost.

   (c) There are no externalities.

      i. An externality is, intuitively, a consequence of trading which affects somebody other than the people engaging in that trade—*e.g.*, if I sell my river access to a factory, and the factory pollutes the river, this will negatively affect those downstream of the river, though they have no say in whether the trade takes place. Since those downstream disprefer the pollution, it is a negative externality of the trade.

      ii. Suppose additionally that the factory plans to offer many jobs to the local workers. This will improve the lives of the local workers. However, they have no say in whether the trade takes place. Since they prefer the availability of jobs, it is a positive externality of the trade.

   (d) Nobody is able to influence prices—there are not monopolies or monopsonies, for instance.

   It is worth stressing that these conditions are not met in the real world, and that, when they are violated, the state of affairs which results can fail to be a Pareto optimum.

3. The first fundamental theorem tells us that what we get from the free market will be a Pareto optimum. Assuming the Pareto Principle, this really just means that it *lacks* a certain *bad-making* feature: the bad-making feature of being worse than another state of affairs. Even if we accept this, however, we might worry that the state of affairs which results from the free market will be worse than some other Pareto optimum we might have reached instead.

4. Here we could appeal to the *second* fundamental theorem of welfare economics.

   SECOND FUNDAMENTAL THEOREM OF WELFARE ECONOMICS
   For every Pareto optimal state of affairs, that state of affairs may be reached by first redistributing wealth, and then letting the free market take over.

   (a) Here, again, I'm using 'free market' in the technical sense introduced above.

   So suppose that you want a particular Pareto optimum—one in which welfare is spread *evenly* among people, and not hoarded in the hands of a few. The second fundamental theorem tells you that this may be achieved by means of the free market. The 2nd fundamental theorem tells you that this, too, may be achieved by means of the free market.

(a) Does this give us a reason to *actually use* the free market to achieve these Pareto optima? Compare: any city in the continental USA can be reached by tricycle. But this is hardly an argument for traveling by tricycle. We should ask what *additional* benefits the free market has over competitor methods of distributing resources. (There are definitely things to be said here; my point is just that the fundamental theorems do not say them.) Also: can a policy-maker actually *know* which Pareto optimum will result from their redistribution? If not, could that give a reason to use alternative methods of resource distribution?

Chapter 6

Beyond the Pareto Principle

6.1 Social Welfare Functions

1. A reminder, the welfare economist thinks that individual's welfare is a function of how well satisfied their preferences are—we called this thesis *preferentism*.

   (a) By 'satisfaction', we do not refer to any mental state; rather, we mean simply that the preferred state obtains.

   (b) The preferences could be *actual* preferences, or they could be *informed and/or idealized* preferences.

2. The welfare economist also thinks that the overall goodness of a state-of-affairs is some function of the welfare of all the individuals in that state-of-affairs—we called this thesis *welfarism*.

3. Thus, the welfare economist is committed to the view we called *preferentist welfarism*.

   Preferentist Welfarism
   The goodness of any given state-of-affairs is determined by the (actual, informed, and/or idealized) preferences of the individuals in that state-of-affairs.

   (a) For, if we are given everyone's individual preference orderings, $\succeq_i$, then we will be in a position to tell just how well satisfied everyone's preferences are. According to preferentism, this is all that we need in order to know every individual's welfare. And according to welfarism, every individual's welfare is all that we need in order to know how good the state-of-affairs is.

4. Saying just that goodness is *determined by* the preferences of the individuals in that state of affairs does not tell us anything at all about *how* it is so determined. We will want to say something about the particular *kind* of function which determines goodness from preferences.

   (a) Let's call that function, whatever it may be, a *social welfare function*—and we'll denote the function with 'G', for 'goodness' (or, alternatively, for 'group').

(b) Here's how G works: you hand it a bunch of individual preference orderings, $\succeq_1, \succeq_2, \ldots, \succeq_N$, and it hands you back some *betterness* ordering '$\succeq_G$'.

$$G : \begin{bmatrix} \succeq_1 \\ \succeq_2 \\ \vdots \\ \succeq_N \end{bmatrix} \rightarrow \succeq_G$$

5. The preferentist welfarist is in need of a social welfare function of this form.

6. Thus far in the course, we've not yet seen any particular social welfare function. However, we've seen a proposed *constraint* on a social welfare function—the Pareto Principle.

   The Pareto Principle
   For all states-of-affairs S, T: if somebody (strictly) prefers state S to state T, and nobody (strictly) prefers state T to state S, then S is better than T.

   $$\text{if, for all individuals } i, S \succeq_i T, \text{ and, for some individual } j, S \succ_j T, \text{ then } S \succ_G T$$

   (a) There are many social welfare functions incompatible with the Pareto Principle. For instance, consider a radically egalitarian social welfare function which says that *equality* is of paramount importance. That is: every unequal states of affairs is worse than every equal state of affairs. Then, suppose that Ike and Janet's utility functions are linear in dollars and consider the following two distributions of money between Ike and Janet:

	Ike's money in s	Janet's money in s
A	\$5	\$5
B	\$10	\$5

   Our proposed radical egalitarian social welfare function tells us that $A \succ_G B$. But the Pareto Principle tells us that $B \succ_G A$.

   (b) The Pareto Principle on its own, however, does not pin down any unique social welfare function. And it is consistent with many obviously immoral social welfare functions.

   (c) For instance, consider a *dictatorial* social welfare function which says that, if *Ike* strictly prefers X to Y, then X is strictly better than Y; and, if Ike is indifferent between X and Y, then Janet's preference can be used to break the tie. This social welfare function abides by the Pareto Principle. (Why?)

6.2 Preferentism and Interpersonal Comparisons

What is the philosophical position of welfarism*? What is the philosophical position of* preferentism*? Why does Hausman think that any welfarist will have to make interpersonal comparisons of welfare? Briefly explain the difference between a* cardinal *and an* ordinal *utility function. Does Hausman think that it is possible for the preferentist welfarist to make interpersonal comparisons of welfare if preferences are measured with cardinal utility functions? If so, how? If not, why not? Supposing that preferences are measured with a cardinal utility function, what is Hausman's objection to welfarist preferentism?*

1. Hausman: the preferentist welfarist needs to supply us with betterness judgments which go beyond the Pareto Principle. (That is: it needs to say more about the social welfare function.) To do this, it must be able to make *interpersonal comparisons* of welfare.

   (a) Distinguish two kinds of interpersonal welfare comparisons: *unit* comparisons and *level* comparisons.

   (b) A *level* comparison says something about whether, in a given state of affairs S, Ike is *better off* than Janet, or whether Janet is *better off* than Ike, or whether Ike and Janet are equally well off. For the preferentist, we can think of this, roughly, as saying whether $\mathcal{U}_i(S) > \mathcal{U}_j(S)$ or whether $\mathcal{U}_i(S) < \mathcal{U}_j(S)$, or whether $\mathcal{U}_i(S) = \mathcal{U}_j(S)$.

   (c) A *unit* comparison says something about whether, in a transition from state S to state T, the amount that Ike's welfare *changed* is greater than or less than the amount that Janet's welfare changed, or whether they both changed by the same amount. For the preferentist, this is, roughly, whether $\mathcal{U}_i(S) - \mathcal{U}_i(T) < \mathcal{U}_j(S) - \mathcal{U}_j(T)$, whether $\mathcal{U}_i(S) - \mathcal{U}_i(T) > \mathcal{U}_j(S) - \mathcal{U}_j(T)$, or whether $\mathcal{U}_i(S) - \mathcal{U}_i(T) = \mathcal{U}_j(S) - \mathcal{U}_j(T)$.

2. Hausman believes that we must be able to make unit comparisons if we are to compare the goodness of policies for which there will be some winners and some losers (some with higher welfare after the policy is enacted than they had before the policy was enacted, and some with lower welfare after the policy is enacted than they had before the policy was enacted). When there are winners and losers, we must be able to say whether the welfare gains of the winners outweigh the welfare losses of the losers. And this requires comparing welfare *differences* in welfare across two individuals. So it requires inter-personal *unit* comparisons.

3. If we think that society has special benefits to the least well off, or that systematically lower welfare among certain groups is *unjust*, e.g., then we will also want to be able to make interpersonal *level* comparisons.

4. However, Hausman does not think that the preferentist welfarist is capable of doing this without making morally unacceptable claims.

   (a) Hausman *does*, by the way, think that (morally acceptable) interpersonal unit and level comparisons of welfare can be made. He just doesn't think that they can be made if preferentism were true. So, he concludes, preferentism must be false. The overall argumentative structure of his paper is this:

   P1. If preferentism is true, then it is not possible to make (morally acceptable) interpersonal comparisons of welfare.

   P2. It is possible to make (morally acceptable) interpersonal comparisons of welfare.

   C. Preferentism is false.

6.2.1 Ordinal Utility Comparisons

5. Firstly, Hausman argues that *ordinal* utility functions do not allow meaningful interpersonal comparisons. That's because, with ordinal utility functions, the only meaningful content is the *order* in which states are ranked.

(a) An attempt: perhaps we should simply count the states-of-affairs above and below the actual outcome in the agent's preference ordering?

i. Hausman's first response: there could be infinitely many states-of-affairs in the preference ordering. Consider the following infinite sequence of states-of-affairs:

$$\ldots, \text{Ike has } \$-2, \text{Ike has } \$-1, \text{Ike has } \$0, \text{Ike has } \$1, \text{Ike has } \$2, \ldots$$

If Ike prefers more money to less, then every possible state has just as many states above it as it has below it: infinitely many.

ii. Hausman's second response: the states over which your preferences are defined depends upon which factors are relevant to you. Suppose that all Ike and Janet care about are the number of tomatoes in the stew, and they know that there are either 2 tomatoes or 3 tomatoes. Ike wants to have two most of all, then one, and then three. So Ike's preference ordering is given by:

$$2 \succ_i 1 \succ_i 3$$

However, Janet doesn't only care about how many tomatoes we eat; she additionally cares about whether there are tomatoes left over. Her preference ordering is given as follows:

$$(2,0) \succ_j (1,1) \succ_j (2,1) \succ_j (1,2) \succ_j (0,2) \succ_j (3,0) \succ_j (0,3)$$

(where '(x, y)' is eating x tomatoes with y left over). Then, if they eat 1 tomato with 1 left over, then Janet will have a higher welfare than Ike does, simply in virtue of the fact that she ranks more outcomes than Ike does.

6.2.2 Ordinal Comparisons via Extended Preferences?

6. Perhaps we could build up a single utility scale for everyone out of "extended" preferences. The things being ranked in this preference ordering wouldn't just be *states*, but rather *states centered on individuals*.

(a) E.g., Ike with cookies while Janet gets cake $\preceq$ Janet with cake while Ike gets cookies.

(b) We could then use these *extended* preferences to build up a utility function $\mathcal{V}(S, i)$ (where S is a state and i an individual) which gives us the utility level of i if the state is S. If S is the state in which Ike gets cookies and Janet gets cake, and if $\mathcal{V}(S, i) \geqslant \mathcal{V}(S, j)$, then Ike has a higher degree of welfare in S than Janet does.

i. Whose preferences are these?

ii. Hausman: they can't be the preferences of other people. For

A. I might prefer to be Socrates dissatisfied than a very satisfied king. This doesn't mean that Socrates' preferences are better satisfied than the king's.

iii. Hausman: they cannot be based upon anybody's mental states, for the satisfaction of preferences is not itself a mental state.

iv. Hausman: they cannot be based upon considerations of well-being, since such preferences are supposed to *ground* claims about well-being; on pain of circularity, we cannot use such judgments to form our preferences.

7. Hausman concludes that there is no way of forming the kinds of preferences required for the extended preferences approach to work.

6.2.3 Cardinal Utility Comparisons

1. Hausman: *there is* a way to measure interpersonal utility when we have *cardinal* utility functions.

   The 'Zero One' Rule
   If Ike and Janet's utility functions are given by $\mathcal{U}_i$ and $\mathcal{U}_j$—*and those utility functions are bounded*—then, to compare Ike and Janet's utilities, give them the arbitrary end points of zero and one by transforming them as follows:

   $$\mathcal{U}_i^{z\text{-}o}(S) = \frac{\mathcal{U}_i(S) - \min \mathcal{U}_i}{\max \mathcal{U}_i - \min \mathcal{U}_i}, \qquad \mathcal{U}_j^{z\text{-}o}(S) = \frac{\mathcal{U}_j(S) - \min \mathcal{U}_j}{\max \mathcal{U}_j - \min \mathcal{U}_j}$$

   Then, say that Ike's utility is higher than Janet's iff $\mathcal{U}_i^{z\text{-}o}(S) > \mathcal{U}_j^{z\text{-}o}(S)$, where S is the actual state (level comparisons), and say that, in a transition from S to T, Ike's utility was raised more than Janet's was if $\mathcal{U}_i^{z\text{-}o}(S) - \mathcal{U}_i^{z\text{-}o}(T) > \mathcal{U}_j^{z\text{-}o}(S) - \mathcal{U}_j^{z\text{-}o}(T)$.

2. Hausman's argument for the zero-one rule (assuming, throughout, that preferentism is correct):

   P1. If Ike and Janet are both given their most preferred option, then they should be at the same level of well-being.

   C1. If B_i is Ike's most preferred option (his *best*), and B_j is Janet's most preferred option (her *best*), then $\mathcal{U}_i(B_i) = \mathcal{U}_j(B_j)$. [from P1]

   P2. If Ike and Janet are both given their least preferred option, then they should be at the same level of well-being.

   C2. If W_i is Ike's least preferred option (his *worst*), and W_j is Janet's least preferred option (her *worst*), then $\mathcal{U}_j(W_i) = \mathcal{U}_j(W_j)$. [from P2]

   C3. The zero one rule is the correct method for making Interpersonal comparisons of utility [from C1 and C2]

   (a) An objection to **P1** and **P2**: Ike and Janet's preference orderings are merely providing *intrapersonal* judgments about well-being. Nothing about interpersonal comparisons follow from these kinds of judgments. That Ike is at the top of his ranking tells you merely that Ike is better than he *would* be with any other outcome, and similarly for Janet. Nothing about their *comparative* levels of well-being follows from this.

      i. Hausman: if we have no way of comparing interpersonal levels of well-being, then this is bad for Preferentism, since it will not allow us to make judgments of the comparative goodness of almost all the difficult choices we need to make. So suppose that we have *some* way of making interpersonal comparisons of well-being level. Then, it should satisfy the following two constraints:

         equal-to-equal
         If Janet's preferences are identical to Ike's, and they are both at the same point in their preference ordering, then their levels of well-being are identical.

         lower-to-lower
         If Janet's comparative level of well-being is lowered, then her non-comparative level of well-being is lowered.

ii. But these principles are enough to establish P1 and P2. We may argue as follows: Suppose, for the purposes of deriving a contradiction, that Janet and Ike are both at the top of their preference orderings, but that Janet is better off than Ike. By EQUAL-TO-EQUAL, were Janet's preferences to become identical with Ike's, her comparative well-being would go down. By LOWER-TO-LOWER, were Janet's preferences to become identical to Ike's, her non-comparative level of well-being would go down. (What about the case where Janet and Ike are at the top of their preference ordering by Janet is *worse* off than Ike?)

iii. But this, Hausman contends, is incompatible with Preferentism. For *the only thing* which determines how well off Janet is is how well her preferences are satisfied. If her preferences remain maximally satisified, then her (non-comparative) level of well-being should not go down. So it is impossible for Janet and Ike to both be at the top of their ordering but still differ in their non-comparative levels of utility.

iv. By similar reasoning, we may also get the conclusion that it is impossible for both Janet and Ike to be at the *bottom* of their preference ordering and still differ in their non-comparative levels of utility.

v. So, we have P1 and P2, from which Hausman has argued the zero-one rule follows.

3. Hausman thinks that the zero one rule is the right way to make interpersonal utility comparisons, but he thinks that, in conjunction with welfarist preferentism, the ethical judgments it yields are implausible. So, he thinks that we should reject welfarist preferentism.

> **P1**. If Ike is satiated more easily than Janet, then welfarist preferentism, with the zero one rule, tells us that it is better to give more to Janet than Ike.
>
> **P2**. It is not better to give more to Janet than Ike, simply because Ike is more easily satiated.
>
> ---
>
> **C**. Welfarist preferentism, with the zero one rule, is false.

4. Let's think about **P1** a bit more: it assumes something about the social welfare function. If, *e.g.*, the social welfare function is just the *sum* of everyone's individual (zero-one) welfare,

$$\mathcal{U}_G(S) = \sum_i \mathcal{U}_i^{z\text{-}o}(S)$$

then **P1** looks very plausible. But the social welfare function *G needn't* just be a straight sum of the zero-one utilities of everyone. Perhaps if we are more careful about how we specify our social welfare function, we can avoid the embarrassing consequence **P1**. Let's start with the Pareto Principle and think about some minimal ways of *extending* the claims about betterness that we get from the Pareto Principle.

6.3 Extending the Pareto Principle

> *Introduce and explain the Kaldor-Hicks Principle. Say how it could be viewed as a justification of Cost-Benefit Analysis, and present the objection to it discussed in class.* For Bonus points*: introduce and explain the Scitovsky Principle, and present the objection to it discussed in class.*

1. Recall: the Pareto Principle does not allow us to make comparisons between Pareto optimal states of affairs (or even between any two states where neither is a Pareto Improvement on the other, even if one is a Pareto optimum and the other is not).

2. Suppose that we don't want to specify a total social welfare function. Perhaps we can get by with a principle *like* the Pareto Principle, but which is slightly more informative than it is.

3. In fact, there is a method for extending the comparisons allowed by the Pareto Principle. This method, which we will refer to as as the Kaldor-Hicks Principle, underlies the appraisals of *cost-benefit analysis*.

   (a) A cost-benefit analysis is used to make comparisons between policies where there are both winners and losers.

   (b) Roughly, what you do in a cost-benefit analysis (CBA) is tally up the total costs, measured in dollars, of the loses incurred by the losers, and you tally up the total costs, again measured in dollars, of the gains of the winners. Measured in this way, if the winners win more than the losers lose, a CBA says the policy is better than the status-quo.

6.3.1 the Kaldor-Hicks Principle

4. To understand the appraisals offered by the Kaldor-Hicks principle, let us first introduce the idea one state being a *redistribution* of another.

   REDISTRIBUTION
   One state of affairs, r, is a *redistribution* of another state of affairs, s, iff, from state s, we can reach state r simply by transferring goods from some individuals to others.

5. For instance, suppose that Ike and Janet's utility functions are linear in dollars, and consider the following states of affairs: a, b, c, d, and e:

s	Ike's money in s	Janet's money in s
a	\$6	\$1
b	\$4	\$4
c	\$5	\$4
d	\$6	\$2
e	\$7.5	\$1.5

   (a) Here, d is a redistribution of b, and e is a redistribution of c. (And, of course, symmetrically, b is a redistribution of d and c is a redistribution of e.) We can get from b to d by taking two dollars from Janet and giving them to Ike. And, likewise, we can get from c to e by taking two and a half dollars from Janet and giving them to Ike.

6. By the way, in this example, both d and e are Pareto improvements on a, and c is a Pareto improvement in b. There are no other Pareto improvements. Since the Pareto Principle tells us that Pareto improvements are better than the things they are Pareto improvements *on*, the Pareto Principle tells

us this, and no more, about the betterness ordering $\succ_G$,

$\succ_G$	a	b	c	d	e
a					
b					
c		✓			
d	✓				
e	✓				

Now, we may want to make some additional assumptions here, like, e.g., that if $x \succ_G y$, then $y \not\succ_G x$ and that $x \not\succ_G x$. If we make these additional assumptions, we can learn this (but no more) from the Pareto Principle:

$\succ_G$	a	b	c	d	e
a	×			×	×
b		×	×		
c		✓	×		
d	✓			×	
e	✓				×

7. The basic idea behind the Kaldor-Hicks principle is to extend the Pareto Principle by considering not just Pareto Improvements, but additionally what we will call *Kaldor-Hicks improvements*. A state x is a Kaldor-Hicks improvement on a state y iff it is possible to redistribute goods in x so that we end up with a state, z, which is a Pareto improvement on y.

   Kaldor-Hicks Improvement
   If there is some state-of-affairs r which is both a redistribution of s and a Pareto improvement on t, then s is a Kaldor-Hicks improvement on t.

   (a) Think of it like this: for many policies, there will be both winners and losers. But suppose that the winners would be able to compensate the losers in such a way that, after this compensation takes place, you end up in a state which is a Pareto improvement upon the state that you started out in. Then, the policy is a Kaldor-Hicks improvement—*even though this compensation never actually occurs*.

   (b) Note that, since every state is a redistribution of itself, any Pareto improvement is a Kaldor-Hicks improvement.

   (c) In the example from above, state b is a Kaldor-Hicks improvement on a—since d is a redistribution of b, and d is a Pareto improvement on a. Similarly, state c is Kaldor-Hicks improvement on a—since e is a redistribution of c, and e is a Pareto improvement on a.

8. With the notion of a Kaldor-Hicks improvement under our belts, we can introduce the Kaldor-Hicks Principle. This principle just says that Kaldor-Hicks improvements are improvements *simpliciter*. If one state is a Kaldor-Hicks improvement on another, then it is *better* than that other state.

   Kaldor-Hicks Principle
   If x is a Kaldor-Hicks improvement on y, then x is better than y.

   (a) Notice that, since every Pareto improvement is a Kaldor-Hicks improvement, the Kaldor-Hicks Principle entails the Pareto Principle.

9. Looking back at the example from before, since b and c are both Kaldor-Hicks improvements on a, the Kaldor-Hicks Principle allows us to glean the following additional information about the betterness relation $\succ_G$:

$\succ_G$	a	b	c	d	e
a					
b	✓				
c	✓	✓			
d	✓				
e	✓				

As before, we may want to make some additional assumptions here, like, e.g., that if $x \succ_G y$, then $y \not\succ_G x$, and that $x \not\succ_G x$. If we make these additional assumptions, we can learn this from the Kaldor-Hicks Principle:

$\succ_G$	a	b	c	d	e
a	×	×	×	×	×
b	✓	×	×		
c	✓	✓	×		
d	✓			×	
e	✓				×

   (a) This is a substantial improvement on the Pareto Principle. We can now definitely say that we should *not* pursue state a. For every other alternative is strictly better than a.

10. Above, we assumed (as is entirely natural) that if $x \succ_G y$, then $y \not\succ_G x$. That is, we assumed that betterness is *asymmetric*. Betterness had better be asymmetric; but there's a substantial worry that, if the Kaldor-Hicks Principle is true, then it will not be.

   (a) Consider figure 6.1. The horizontal axis is Ike's utility, and the vertical axis is Janet's utility. Suppose that there are two policies we could pursue: *e.g.*, we will either build a subway near Ike or near Janet. If we build the subway near Janet, then we'll end up at state b, and the possible redistributions will lie on the curve running through b and f. If we build the subway near Ike, then we'll end up at state c, and the possible redistributions will lie on the curve running through c and e. Then, e is a redistribution of c and f is a redistribution of b. Since e is a Pareto improvement on b, the Kaldor-Hicks Principle tells us that

$$c \succ_G b$$

   And, since f is a Pareto improvement on c, the Kaldor-Hicks Principle tells us that

$$b \succ_G c$$

   So the Kaldor-Hicks Principle tells us that betterness is not asymmetric.

11. This is a serious problem; it provides the following, incredibly forceful argument against the Kaldor-Hicks Principle.

   P1. Betterness is an asymmetric relation.

   P2. If the Kaldor-Hicks Principle is true, then betterness is not an asymmetric relation.

   C. The Kaldor-Hicks Principle is not true.

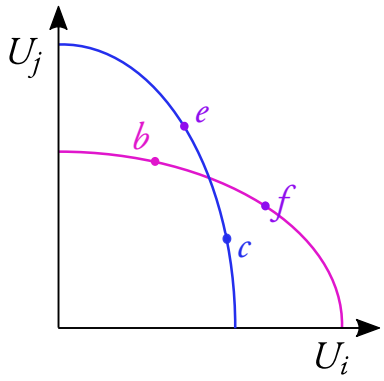


Figure 6.1: The Kaldor-Hicks Principle leads to violations of the asymmetry of betterness.

6.3.2 The Scitovsky Principle

12. But perhaps we can get around these issues with asymmetry. Here's one suggestion: we simply rule out of hand those cases in which there are asymmetry violations. That is, if x is a Kaldor-Hicks improvement on y, and also y is a Kaldor-Hicks improvement on x, then, in that case, we *won't* say that x is better than y.

13. Here's a way of implementing this approach: we'll say that x is a *Scitovsky* improvement on y just in case x is a Kaldor-Hicks improvement on y, and y is *not* a Kaldor-Hicks improvement on x.

    Scitovsky Improvement
    If x is a Kaldor-Hicks improvement on y, *and additionally* y is *not* a Kaldor-Hicks improvement on x, then x is a *Scitovsky improvement* on y

    (a) Thus, in figure 6.1, neither b nor c is a Scitovsky improvement on the other.

    (b) Notice that, if x is a Scitovsky improvement on y, then x is automatically a Kaldor-Hicks Improvement on y.

    (c) Notice also that, since every Pareto improvement is automatically a Kaldor Hicks improvement; and since, if x is a Pareto improvement on y, then y cannot be a Pareto improvement on x, it follows that, if x is a Pareto improvement on y, then x is a Scitovsky improvement on y.

    (d) So there is the following logical relationship amongst Pareto improvements, Scitovsky improvements, and Kaldor-Hicks improvements:

    $$\text{Pareto improvement} \Rightarrow \text{Scitovsky improvement} \Rightarrow \text{Kaldor-Hicks improvement}$$

14. And we can then endorse what we'll call the *Scitovsky Principle*, which says that Scitovsky improvements are improvements *simpliciter*.

    Scitovsky Principle
    If x is a Scitovsky improvement on y, then x is better than y.

    (a) Because of the logical relations between the various kinds of improvements, the Scitovsky Principle is intermediate in strength between the Pareto and the Kaldor-Hicks Principle.

    $$\text{Pareto Principle} \Rightarrow \text{Scitovsky Principle} \Rightarrow \text{Kaldor-Hicks Principle}$$

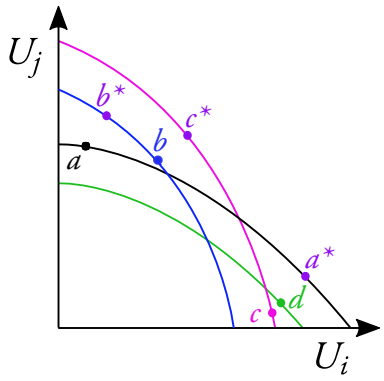


Figure 6.2: The Scitovsky Principle leads to cycles of betterness

15. However, there are problems with the Scitovsky Principle, too.

    (a) Consider figure 6.2. There, the horizontal axis is Ike's utility, and the vertical axis is Janet's utility.

    (b) Because b^* is a redistribution of b, and b^* is a Pareto improvement on a, b is a Kaldor-Hicks improvement on a.

    (c) And no redistribution of a is a Pareto improvement on b, so b is a Scitovsky improvement on a. Wherefore the Scitovsky principle tells us that

    $$b \succ_G a$$

    (d) Similarly, c^* is a redistribution of c, and c^* is a Pareto improvement on b. So c is a Kaldor-Hicks improvement on b.

    (e) And no redistribution of b is a Pareto improvement on c, so c is a Scitovsky improvement on b. Wherefore the Scitovsky principle tells us that

    $$c \succ_G b$$

    (f) d is a Pareto improvement on c, and any Pareto improvement is a Scitovsky improvement. So the Scitovsky principle tells us that

    $$d \succ_G c$$

    (g) Finally, a^* is a redistribution of a, and a is a Pareto improvement on d. So a is a Kaldor-Hicks improvement on d.

    (h) Additionally, there is no redistribution of d which is a Pareto improvement on a, so a is a Scitovsky improvement on d. Wherefore the Scitovsky principle tells us that

    $$a \succ_G d$$

16. But look at what we've now concluded. We've now concluded each of the following:

$$b \succ_G a$$
$$c \succ_G b$$
$$d \succ_G c$$
$$a \succ_G d$$

But this can't be. *Betterness* is an acyclic relation—there are no cycles of betterness.

17. This is a serious problem. It affords the following very powerful argument against the Scitovsky Principle:

    P1. Betterness is an acyclic relation.

    P2. If the Scitovsky Principle is true, then betterness is not an acyclic relation.

    C. The Scitovsky Principle is not true.

Chapter 7

Wrongful Exploitation

7.1 Welfare Economics and the Market

1. Recall, the theory of welfare economics consists of three claims—one about welfare, one about goodness, and one about rightness:

   PREFERENTISM
   Your welfare is strictly higher in state-of-affairs S than in state-of-affairs T if and only if you (actually/informedly/ideally) prefer S to T.

   WELFARISM
   Whether a state-of-affairs S is better than a state of affairs T is determined entirely by the welfare of the persons in those states-of-affairs.

   CONSEQUENTIALISM
   Whether an act is morally *right* is determined by the goodness of the state-of-affairs which results (or is expected to result) from the act's performance.

2. Preferentist welfarism (the first two claims of welfare economics) seems to commit us to at least the following constraint on a betterness ordering $\succ_G$:

   PARETO PRINCIPLE
   If somebody strictly prefers state-of-affairs S to state-of-affairs T, and nobody strictly prefers T to S, then S is better than T,

   $$\text{if, for some } i,\ S \succ_i T \ \text{ and, for all } j,\ T \not\succ_j S, \ \text{ then } \ S \succ_G T$$

   (a) The Pareto Principle is a claim about *axiology*—a claim about which things are valuable. It says that Pareto improvements are genuine improvements; they make the world a better place.

3. As we have seen, the Pareto Principle is a very weak axiological claim. However, it can be used to offer an argument for using the free market to distribute goods (after, perhaps, some redistribution). These are the fundamental theorems of welfare economics.

   FIRST FUNDAMENTAL THEOREM OF WELFARE ECONOMICS
   The state of affairs resulting from a *free market* will be Pareto Optimal.

SECOND FUNDAMENTAL THEOREM OF WELFARE ECONOMICS
For every Pareto optimal state of affairs, that state of affairs may be reached by first redistributing wealth, and then letting the *free market* take over.

(a) In our statement of these theorems, we use 'free market' in a slightly technical sense. We assume two things in particular (we assume more besides, but these assumptions will be relevant today):

i. People's utilities/preferences are not interdependent—I don't want you to get what you want (nor do I want you to *not* get what you want), and you do not want me to get what I want (nor do you want me to *not* get what I want).

ii. There are no—positive or negative—externalities. That is: goods are owned and enjoyed by one and only one person.

7.2 Higher Values, the Market, and Degradation

According to Elizabeth Anderson, which kinds of goods are appropriately treated as commodities, and why? Provide an example of a good which she does not think are appropriately treated as commodities, and explain how this follows from what you've said before. Then, explain when Anderson thinks goods can be debased or degraded, and provide an example of a good being degraded through commodification.

1. Elizabeth Anderson: *commodities* are objects which are valued in a certain way—they are valued according to the mode of *use*.

   (a) To *merely use* something is to recognize only the *instrumental* value it has *for you*, and to not recognize any *intrinsic* value it may have, or the value it may have *to others*.

      i. *E.g.*, some owners of David Smith's sculptures removed the paint from their sculptures because the unpainted works were selling for higher value than the painted works. These people were *merely using* the sculptures—treating them as only instrumentally valuable, and as only having instrumental value *for them*.

2. Anderson believes that what the fundamental theorems of welfare economics show us is that the market is an effective means for allocating items which merely have *use value*. However, she does not think that everything which has value merely has it because it satisfies somebody's preferences; she does not think that use value is the only kind of value there is. She believes, rather, that some things have non-instrumental value; and some things have value for others besides yourself.

   (a) E.g., historical architecture may have *shared* aesthetic value for people besides the owner. (It may also have *intrinsic* value, as a thing of beauty.)

   (b) E.g., human life (or perhaps animal life, too) may have *intrinsic* value, quite apart from the instrumental, use value (other) humans can derive from it.

3. Because she recognizes other values like these, Anderson does not believe that everything is appropriately treated as a commodity. Consider a few examples:

   (a) The British government 'privatizes' Stonehenge, and auctions it off to the highest bidder. McDonald's purchases it and begins selling burgers there.

(b) The U.S. government opens a market for votes. Buyers are able to purchase the votes of sellers and use them at the ballot box.

(c) A company begins selling friendship. Employees, most of them poor, are paid by the agency to befriend their, mostly rich, clientele.

In cases like these, Anderson believes that something of intrinsic or shared value has been *debased* by subjecting it to the norms of the market. Not all goods, then, are appropriately treated as commodities; not all goods are to be traded in an open market.

4. Which things are appropriately treated as commodities?

   (a) Anderson: to answer this, we must consider which kinds of values the market is in a position to successfully appreciate.

   (b) There are 5 important features of the system of norms embodied in market transactions:

      1. Market relations are impersonal
         i. Market relationships are relationships between strangers who are free to abandon their business relationship at any time. They are unlike, *e.g.*, familial relationships or friendships.
      2. Within the market, you are free to pursue your own personal advantage without consideration for the advantage of others.
         i. This is, Anderson stresses, one of the reasons that the market is so effective. The "invisible hand" guaranteed by the first fundamental theorem depends upon individuals only caring for their own self-interest. The mechanism would break down if traders cared for the well being of their trading partners.
      3. Good traded on the market are exclusive and rival in consumption.
         ▷ A good is *exclusive* if the seller is able to restrict access to the good; and only provide it to those who pay.
         ▷ A good is *rival* if the amount that one person consumes decreases the amount that others are able to consume. (As opposed to, *e.g.*, a joke or a piece of knowledge.)
      4. Within the market, all matters of value are simply matters of personal taste.
         ▷ This means in particular that the market does not distinguish between urgent needs and intense desires.
      5. Dissatisfaction with a market relationship is expressed by "exit" rather than "voice".
         ▷ If I am unhappy with the terms of your offer, then my only option is to reject the offer. I may not express my views about how the good will be designed or marketed.

5. Anderson: the goods which ought to be treated like commodities are those whose "production, distribution, and enjoyment is properly governed by these five norms, and its value may be fully realized through use."

   Conversely, those goods whose production, distribution, and enjoyment is not properly governed by these five norms should *not* be commodified.

   (a) For instance, personal relationships are properly governed by a different system of norms.

- ▷ Personal relationships are not *im*personal; you are not free to solely pursue your own personal advantage; the good of a personal relationship is not rival; and dissatisfaction with a personal relationship is properly expressed through "voice".
- ▷ While the good of a market relationships may be fully realized with mutually beneficial *trades*, the good of a personal relationships may only be fully realized with *gifts*.
  - Gifts are not given as a *quid pro quo*, with an expectation of repayment, and they are meant to symbolize something about the nature of the relationship.
  - This is why Anderson thinks it is bad to give cash as a gift, and why it is upsetting and unsettling for a friend to be too anxious to "settle accounts". It makes it appear as though the relationship is an impersonal market relationship rather than a personal relationship.

6. When goods which are properly valued through personal relationships become commodities—when they are valued under the mode of *use* rather than valued as *gifts*—Anderson believes that those goods are *debased* or *degraded*. More generally, a good is debased when it is treated in accordance with a lower mode of valuation than is proper to it. Some examples:

   (a) Prostitution debases the good of the personal sexual relationship, which is properly be treated as a gift, not as an exchange.

   (b) A McDonald's at Stonehenge debases the intrinsic and shared aesthetic value of that historical site.

   (c) Selling and buying votes debases the democratic value of the vote—an inalienable vote expresses fraternal relations of treating each individual as an equal in collective deliberation.

7. Not only are the commodified *goods* debased through commodification, but the one who sells goods properly valued as gifts is similarly debased.

   (a) Individuals are debased or degraded when *they* are treated in accordance with a lower mode of valuation than is proper to *them*.

   (b) Thus, the prostitute is debased by providing their sexuality—something which should be properly provided as a gift—as a commodity.

   (c) So too would a 'friendship worker' be debased by providing their friendship as a commodity.

8. Anderson thinks that a particular kind of *wrongful exploitation* can occur when one side of a relationship is providing goods according to the norms of gift exchange while the other side is returning them according to the norms of market transaction.

   (a) E.g., when corporations attempt to establish familial relationships with employees. The corporation is, Anderson believes, attempting to get their employees to provide their labor as a gift to the company, while the company continues to provide wages under the norms of market transaction.

   (b) In such situations, the corporation is *wrongfully exploiting* the workers by misleading them about the nature of the relationship in order to extract more goods from their employees than they provide in return.

9. What is it for one person to *wrongfully exploit* another?

7.3 Price Gouging and Sweatshop Labor

> *What is Zwolinski's argument that Price Gouging and Sweat Shop labor is not wrongfully exploitative? Introduce and explain* either *Valdman's* or *Christiano's theory of wrongful exploitation (and say which it is). What does this theory say about the case of Price Gouging and Sweat Shop Labor?*

1. 'Price gouging' occurs when a vendor substantially raises the prices of important goods in the aftermath of a natural disaster.
   (a) E.g., raising the price of water to exorbitant levels after a hurricane.
2. Here is an argument, from Zwolinski, that we should not have laws against price gouging:

   P1. In standard cases of price gouging, some sellers would not offer goods at a price below the 'gouging' one.

   P2. 'Price gouging' trades are mutually preferred to the lack of trade.

   P3. In standard cases of price gouging, neither buyer nor seller lack any relevant information.

   P4. When neither buyer nor seller lack any relevant information, mutually preferred trades make both parties better off.

   C1. In standard cases of price gouging, laws against price gouging would prevent trades which leave both parties better off.

   P5. Laws which will, in the standard case, leave people worse off are wrong.

   C2. Laws against price gouging are wrong.

   (a) Notice that (P4) is expressing a form of preferentism; and (P5) is expressing a kind of welfarist consequentialism.
   (b) A note on this argument: you might worry that, even though laws against price gouging will prevent *some* mutually-preferred trades from taking place, it will nevertheless allow *other* mutually-preferred trades to take place at a lower price. Those who engage in these trades will prefer the state-of-affairs in which the price-gouging law is in effect. So, we should note, the lack of price gouging laws is not a Pareto improvement over the presence of a price gouging law.
      - ▷ Taking note of this subtlety, we might interrogate (P5) a bit further. Is it saying that laws which will make *some* people worse off are wrong, even if they leave *others* better off? In this case, we may wish to reject the premise. Or, alternatively, is it saying that laws which leave *everybody* worse off are wrong? In that case, we may wish to object that the conclusion (C2) does not follow from (C1) and (P5).
      - ▷ Getting around this objection will require us to go beyond the Pareto Principle and to make some kinds of interpersonal comparisons of well-being.

3. There are other, similar cases to price gouging which raise similar issues. For instance, sweat shop labor is mutually preferrable to its absence—as evidenced by the fact that laborers agree to work in these conditions rather than not. If we suppose that laws against sweatshop conditions would lead to these jobs disappearing, then we will have an argument that laws against sweat shops are wrong which parallels the argument above.

    P1′. In standard cases of sweat shops, some employers would not offer jobs at conditions above the 'sweat shop' level.

    P2′. Sweat shop employment is mutually preferred to the lack of employment.

    P3′. In standard cases of sweat shops, neither employer nor employee lack any relevant information.

    P4′. When neither employer nor employee lack any relevant information, mutually preferred trades make both parties better off.

    C1′. In standard cases of sweat shops, laws against sweat shops would prevent trades which leave both parties better off.

    P5′. Laws which will, in the standard case, leave people worse off are wrong.

    C2′. Laws against sweat shops are wrong.

    (a) Here, again, we might note the subtlety with (P5′).

4. In response to both arguments, some wish to object to the fifth premise on the following grounds:

    (a) Some mutually preferred trades are *wrongfully exploitative*. (Wrongful exploitation is a kind of *unfair advantage taking*.) Laws preventing such wrongful exploitation can be required, even if they prevent mutually preferred trades.

    (b) Notice that, if we favor laws against price gouging because they are wrongfully exploitative—and we accept preferentist welfarism—then we will be denying *consequentialism*. We will think that laws can be morally right, even if they lead to worse states-of-affairs.

5. Zwolinski responds to these kinds of worries by denying that mutually preferred trades are wrongfully exploitative—or, more exactly, that this follows from typical committments that those objecting to price gouging and sweat shops typically accept. He offers the following argument:

P6. If it is permissible for A to not trade with B, and trading with B is better for B than not trading at all, then it is not wrong for A to trade with B.

P7. It is permissible for venders (employers) to not sell their goods after a natural disaster (to not offer employment in less developed countries).

P8. Selling goods at inflated prices is better for disaster victims than not selling them goods at all (offering jobs in substandard conditions is better for employees than not offering any jobs at all).

C3. It is not wrong for vendors to sell their goods at exorbitant prices after natural disasters (for employers to employ people in sweat shop conditions).

C4. It is not wrongfully exploitative for vendors to sell their goods at inflated prices (for employers to employ people in sweat shop conditions).

7.4 Theories of Wrongful Exploitation

6. To begin to think about what wrongful exploitation is, and whether it should lead us to condemn price gougers/sweat shop employers and/or support laws banning price gouging/sweat shop employment (*e.g.*), philosophers have attempted to formulate general *theories*—what I'll sometimes call an *account*—of when a person, A, wrongfully exploits—or takes unfair advantage of—another person, B.

   (a) A theory like this will take the following form (we treat the names 'A' and 'B' as placeholders, for which any two people or groups of people may be substituted):

   A wrongfully exploits B if and only if _____.

   (b) Such a theory should really be separated out into two separate components:

   (**Suff**) A wrongfully exploits B if _____.

   (**Nec**) A wrongfully exploits B only if _____.

   (c) The 'if' part of the theory gives a *sufficient condition* for wrongful exploitation. The 'only if' part gives a *necessary condition* for wrongful exploitation.

   (d) An objection to a theory of this sort could either object to the sufficiency part of the theory or the necessity part of the theory.

7. Here are some relatively uncontroversial cases of wrongful exploitation, or unfair advantage taking (in all cases, 'A' is the name of the exploiter, and 'B' is the name of the exploited):

   Antidote
   B is bitten by a snake. A possess an antidote which sells for $10. B will die if they do not receive this antidote from A. A offers B the antidote for $20,000. B agrees.

   Rescue
   B is drowning. A offers to throw B a life saver in exchange for B's entire life savings. B agrees.

In each of these cases, there is a trade—$20,000 in exchange for the antidote, B's life savings in exchange for the life saver. Note that both parties prefer this trade taking place to the trade not taking place (though, of course, B would prefer the trade taking place at a lower price).

8. Here are two reasons you may think A has done something wrong in these cases:

   (a) B was not in the right kind of position to make the trade—perhaps because B's participation wasn't voluntary, or perhaps because they weren't in a position to refuse. In Christiano's terminology: there was something *procedurally* wrong with the exchange.

   (b) The goods exchanged were not equal in value—the antidote was not worth $20,000, and the throwing of the life saver was not worth B's entire life savings. In Christiano's terminology: there was something *substantively* unequal about the exchange.

9. A schematic procedural account:

   A wrongfully exploits B if and only if A benefits from an exchange with B when B was not participating voluntarily.

   (a) A question: what is it for A to not participate voluntarily in the trade? Perhaps: B cannot reasonably refuse A's offer. B has an urgent need for the antidote or the life saver, and A has a monopoly on these goods.

   Then, our procedural account becomes:

   Procedural A wrongfully exploits B if and only if A benefits from an exchange with B when B cannot reasonably refuse the exchange.

   (a) Christiano: consider the following cases:

      Cheap Antidote
      B is bitten by a snake. A possess an antidote which sells for $10. B will die if they do not receice this antidote from A. A offers B the antidote for $10. B agrees.

      Cheap Rescue
      B is drowning. A offers to throw B a life saver if B pays to replace it. B agrees.

      Since B cannot reasonably refuse A's offer, **Procedural** says that, in these cases, too, A is taking unfair advantage of B. But this is incorrect. A does not wrongfully exploit B in these cases. So **Procedural** is incorrect.

      ▷ Question: does this show that benefitting from a trade with B when B cannot reasonably refuse the exchange is not *sufficient* for exploitation—or does it show that it is not *necessary*?

10. A schematic substantive account:

    A wrongfully exploits B if and only if A *excessively* benefits from an exchange with B.

    (a) A question: what makes a benefit *excessive*?

    (b) *Not*: A's utility for the trade greatly exceeds B's utility for the trade. If this is how we thought of excessive benefits, then we would say that, were A to sell B the antidote for $10, B would be wrongfully exploiting A. (For B *very strongly* desires the antidote, and A only *very weakly* desires the $10.)

(c) Perhaps: A's benefits are excessive iff they are much greater than the *market* price.

Then, our substantive account becomes:

Substantive A wrongfully exploits B if and only if A secures a benefit from their exchange with B which is much greater than the market price.

(a) Christiano: it seems that market prices can *themselves* be exploitative. If the labor market is flooded with workers, the market wage may be pennies a day, but such wages are still wrongfully exploitative—they still take unfair advantage of the worker's desperate condition.

▷ Question: does this show that securing a benefit above the market price is not *sufficient* for exploitation—or does it show that it is not *necessary*?

(b) Christiano: consider the following case:

> Car
> B goes to buy a car from A. A expects to negotiate, and so begins by offering the car at a price much higher than the market value. B is wealthy enough that they do not care to haggle, and they accept A's first offer.

In this case, **Substantive** says that A has wrongfully exploited B. But this is incorrect. A has not taken unfair advantage of B.

▷ Question: does this show that securing a benefit above the market price is not *sufficient* for exploitation—or does it show that it is not *necessary*?

11. Valdman suggests a *hybrid* theory of wrongful exploitation, which blends aspects of procedural and substantive accounts. In schematic form:

A wrongfully exploits B if and only if A benefits *excessively* from an exchange with B when B cannot reasonably refuse the exchange.

(a) Question: what do we mean by 'excessive'? Valdman suggests: A benefits *excessively* from their exchange with B iff they secure much greater benefits than they *would* be able to secure *were* B able to reasonably refuse the exchange—if, for instance, A did not hold a monopoly.[1]

Then, Valdman's hybrid account becomes:

Valdman A wrongfully exploits B if and only if A benefits from an exchange with B when B cannot reasonably refuse the exchange, and A secures much greater benefits than they would be able to secure, *were* B able to reasonably refuse the exchange.

(a) Christiano: consider the following case:

> Lying Surgeon
> B has a disease but does not wish to undergo surgery, if at all possible. A (a suregeon) lies to B, telling them that surgery is the only treatment—when, in fact, there is a non-surgical alternative. A offers a fair, market price for the surgery, which B accepts.

[1] I'm ignoring Valdman's distinction between 'securing' and 'extracting' a benefit. See Valdman, M. 2009. *A Theory of Wrongful Exploitation*. Philosophers' Imprint. vol. 9 (6).

In this case, **Hybrid** says that A has not wrongfully exploited B. But this is incorrect. A has taken unfair advantage of B.

▷ Question: does this show that securing excessive benefits when B cannot reasonably refuse is not *sufficient* for exploitation—or does it show that it is not *necessary*?

12. Against all of these accounts, Christiano offers the following theory of wrongful exploitation built around the idea that exploitation involves violating duties.

Christiano A wrongfully exploits B if and only if A benefits from an exchange with B by violating a duty A has to B.

Christiano believes that this deals with the cases above in the following way:

(a) In Antidote and Rescue: A has a duty to save B's life (for a reasonable price). 'Hard' bargaining signals to B that A will not save B unless B pays more than a reasonable price. So A takes unfair advantage of B.

(b) In Cheap Antidote and Cheap Rescue: 'soft' bargaining is only used to secure a reasonable price, and so, in bargaining, A does not violate their duty to save B's life for a reasonable price. So A does not take unfair advantage of B.

(c) In Car, A does not have a duty to sell B a car at the market price. So A does not take unfair advantage of B.

(d) In Lying Surgeon, A has a duty to not lie to B. So A takes unfair advantage of B.

13. Valdman offers a counterexample to any theory, like Christiano's, which appeals to people's duties or obligations:

> Rescue for Hire
> B is drowning in dangerous waters. The waters are dangerous enough that A has no duty to save B. Nevertheless, A offers to save B in exchange for B's entire life savings. B agrees.

Valdman: in this case, A is taking unfair advantage of B. But A is not violating a duty to B, since A has no such duty.

14. Given Valdman's and Christiano's theories, some questions:

(a) Anderson believed that companies exploit workers by establish familial relationships, so that workers give labor as a gift, while the company treats their labor on the model of a trade, governed by market norms. Does Valdman's theory vindicate Anderson? Does Christiano's theory?

(b) Are sweat shops and price gouging wrongfully exploitative according to Valdman? According to Christiano?

(c) If not, then Zwolinski offered the following argument against their wrongness:

P6. If it is permissible for A to not trade with B, and trading with B is better for B than not trading at all, then it is not wrong for A to trade with B.

P7. It is permissible for venders (employers) to not sell their goods after a natural disaster (to not offer employment in less developed countries).

P8. Selling goods at inflated prices is better for disaster victims than not selling them goods at all (offering jobs in substandard conditions is better for employees than not offering any jobs at all).

C3. It is not wrong for vendors to sell their goods at exorbitant prices after natural disasters (for employers to employ people in sweat shop conditions).

C4. It is not wrongfully exploitative for vendors to sell their goods at inflated prices (for employers to employ people in sweat shop conditions).

Does either theory provide us with any resources to respond to this argument?

Chapter 8

Distributive Justice: Libertarianism & Egalitarianism

8.1 Justice

1. "Justice" is a sub-branch of ethics which concerns our duties (that is: our obligations) to one another. That is, when we investigate *justice*, we are investigating *what we owe to each other*. This excludes, *e.g.*, duties to the environment (if we have such duties), as well as duties to ourselves and duties which are not owed to particular individuals (these are called *impersonal duties*).

   (a) For instance, you may believe that we all have a duty to give money to charity, but that there is no *particular* charity that we owe our money. If so, then this is an impersonal duty, and does not fall within the domain of *justice*.

2. Justice additionally only concerns itself with duties which are *enforceable*. A duty is enforceable iff it is permissible to use force to get people to perform their duties or to use force to punish those who fail to perform their duties.

3. *Distributive* justice concerns the more limited realm of enforceable duties we have to other individuals to provide them with some share of societal resources (duties we may have to *distribute* goods amongst members of a society in one way or another).

8.2 Libertarianism and Rights

> *Explain the distinction between* positive *liberty and* negative *liberty. Say which kind of liberty the libertarian typically focuses on. Then, Explain Hohfeld's four-fold distinction of liberty-rights, claim-rights, power-rights, and immunity-rights; and give an example of each. Finally, briefly explain Nozick's version of Libertarianism. In particular, explain his views about rights as 'absolute side constraints', say what he thinks justice consists in, when he thinks you may justly come to own property, and when you can justly transfer property.*

1. Hausman, McPherson, & Satz: many economists are drawn to a kind of political libertarianism—a political theory of justice focused on the protection of individual liberty. Those attracted to libertarianism might reject, for instance, paternalistic laws like prohibitions on cigarettes or laws requiring you to wear seat belts or helmets.

   (a) A law is *paternalistic* if it interferes with your freedom to act in a certain way, with the intent of improving your overall well-being (*e.g.*, seat belt laws, helmet laws, and prohibition).

   (b) Anti-paternalism is the view that the state should not pass paternalistic laws.

2. Anti-paternalism is supported by the ethical assumptions of welfare economics (assuming *actual*, and not *idealized* preferentism). On those assumptions, people will always choose what they most prefer, and since what people prefer is what's best for them, it is always best to allow people to choose for themselves. Limiting people's choices in the way that a paternalistic law does will make people worse off.

   (a) Actually, the standard assumptions of welfare economics do more than merely advise against paternalistic laws. They entail the impossibility that such laws actually improving anybody's well being. According to actual preferentism, it is impossible for a seat belt law, or an anti-smoking bill, to actually make anybody better off.

   (b) As we've seen, this appears implausible. Seat belt laws and anti-smoking laws appear like they *can* make people better off—this was precisely the reason we rejected *actual* preferentism in favor of *idealized* preferentism. However, it doesn't follow that paternalistic laws are permissible. We could object to paternalistic laws on the grounds that they infringe upon people's *liberty*.

8.2.1 On Liberty

3. What is freedom, or liberty?

   (a) Hausman, McPherson, & Satz: liberty/freedom is a three place relation between a person, P, an objective or *goal*, G, and a potential obstruction, O.

   (b) We could summarize this relation with the English sentence "P is free from O to attain G", or "P is free to achieve G without O". This relation obtains when the potential obstruction O to P's realizing their goal G is not present.

      i. For instance: in America, you are free to speak your mind without governmental punishment. And, in this class, you are free to speak your mind without any grade reduction. In general, however, you are *not* free from the lack of financial and professional means to speak your opinion to a large audience.

      ii. For another: if you are too poor to afford adequate medical insurance, then you are not free from financial impediments to acquire healthcare; however, you *are* free from governmental prohibition to acquire healthcare.

4. My *choice set* is the set of all options which are available to me as objects of choice—it is partially determined by the dimensions along which I am free.

   (a) For instance, if I am free to speak my mind without governmental interference, then my choice set includes the option of speaking my mind and facing no serious consequences. (If not, then it doesn't.)

5. Distinguish three different kinds of liberty:

   (a) Your *effective liberty* is determined by your choice set: the options you have available to you. If your choice set is larger, then you have more effective liberty.

      i. One way of promoting effective liberty: give Rachel the financial means to purchase health insurance, thereby expanding her choice set. (This will likely involve *constricting* the choice set of others through taxation.)

   (b) *Negative liberty* (as we'll use the term) is the liberty to achieve goals without *externally imposed obstructions*.

      i. E.g., I have the negative liberty to read or write any book *without anybody else punishing me*.

      ii. You have the negative liberty to trade with me *without anybody else preventing the trade*.

      iii. You have the negative liberty to engage in any act of consensual sex you wish *without anybody not involved in the act interfering*.

      iv. Perhaps negative liberty is not liberty enough. If you are enslaved by a benevolent master who allows you to do whatever you wish, then you have as much *negative* liberty as anybody else. Because your master is benevolent, you are free from externally imposed obstructions. But your ability to achieve your goals without external interference is *fragile*—it depends upon the arbitrary choice of your master.

         A. Perhaps we should care, not just about negative liberty, but also about *Republican liberty* (note: this has nothing to do with the political party). You have Republican freedom just in case your negative freedom is *guaranteed* by a suitably robust social institution.

   (c) *Positive liberty* is the liberty to be the author of your own actions—it is the liberty you have when your actions are *self-determined*. (This notion is closely related to the Kantian notion of *autonomy*.)

      i. E.g., if you suffer from an addiction to cigarettes, you are free from any external interference, but there is another important sense in which your decision to smoke is not truly *free*. One diagnosis: you are not the *author* of the decision to not smoke; the decision to smoke is not *self-determined*.

      ii. E.g., a brainwashed member of a cult may be free from external impediments to their leaving the cult, but there is another sense in which they are nevertheless not free; they are not the *author* of their decision to remain in the cult.

6. This way of drawing the distinction between negative and positive liberty is Isaiah Berlin's. Perhaps it is wrong, however. Perhaps the right way of drawing the distinction we're interested in is this: *negative* liberty is freedom from *external* obstructions, while *positive* liberty is freedom from *internal* obstructions. (*Cf.* Hausman, McPherson, *&* Satz)

7. Libertarians typically believe that we have the right to certain *negative* liberties which the state should protect; but they deny that we have rights to other forms of liberty. In particular, they deny that we have the right to *positive* liberty, and they are opposed to governmental efforts to promote positive liberty insofar as those efforts infringe upon negative liberties.

   (a) The promotion of positive liberty may consist in paternalistic policies to improve self-determination at the cost of restricting individual choice.

(b) Berlin argues for this position by pointing out that such laws are rife for abuse—authoritarians may claim that certain elites have better knowledge of what will help individuals attain self-determination; and that those elites are therefore justified in all kinds of abuses.

(c) Mill argues against paternalistic policies, not on the ground that they will not promote well-being, but rather on the grounds that policy makers are never in a good enough position to *know* whether they do. The person best situated to evaluate what is best for a person is that very person themselves.

8.2.2 Rights

8. Distinguish *legal* rights from *moral* rights.

   (a) Legal rights are rights which are enshrined in law. Moral rights are independent of any particular legal framework. A system of legal rights can be judged by asking whether it recognizes the moral rights. If you believe that everyone has a right to free speech, then you will be opposed to legal systems which fail to recognize this right.

   (b) One way of making sense of the distinction between legal and moral rights (but not the only way) is this: you have a moral right R iff a system of legal rights *ought* to include R. (For instance: you have a moral right to healthcare iff a system of legal rights ought to include a right to healthcare; you have a right to free speech iff a system of legal rights ought to include a right to free speech.)

   (c) Another way of making sense of the distinction between legal and moral rights is by appealing to Hohfeld's understanding of rights in terms of obligations or duties. Then, moral rights concern *moral* obligations, whereas legal rights concern *legal* obligations.

9. According to Wesley Hohfeld, "first-order" moral rights can be distinguished into two kinds, ultimately definied in terms of duties, or obligations:

   Liberty-Rights

   You have a liberty-right to A iff you do not have an obligation to not A.

   (a) E.g., part of what it is for you to have a right to free speech is that you have a liberty-right to speak freely—*i.e.*, you have no obligation to not speak your mind.

   (b) Relatively, part of what it is for you to have a right to freedom of religion is that you have a liberty-right to practice whichever religion you wish.

   Claim-Rights

   You have a claim-right on some person P, that they A, iff P has a duty (to you) to A.

   (a) E.g., if you have a claim-right to compensation from your employer, then your employer has a duty (to you) to compensate you for your labor.

10. Hohfeld believes that these "first-order" moral rights can sometimes be changed. For instance, my employer can renegotiate the terms of my contract, and thereby alter my claim-right on them. When it comes to such changes, there are similarly rights involved—the right to change the first-order moral rights and the right to have the first-order moral rights *not* be changed in this way. Hohfeld calls these "second-order" moral rights, and distinguishes two kinds:

Power-Rights
You have a power-right with respect to some person P iff you have the ability to alter the first-order rights of P. [Note: here, P could be yourself]

(a) E.g., when you promise to attend your friend's wedding, you confer on your friend a claim-right on you, that you attend their wedding. At the same time, you forego your own liberty-right to not attend the wedding.

(b) When you invite your neighbor into your home, you waive your own claim-right that your neighbor not enter your home. At the same time, you confer on your neighbor a liberty-right to enter your home.

Immunity-Rights
You have an immunity-right with respect to some person P and some first-order right R iff P does not have a power-right to change your first-order right R.

(a) E.g., part of what your right to free speech consists in is an immunity right against others depriving you of your liberty-right to speak your mind.

(b) Relatedly, part of what your right to religious freedom consists in an immunity-right against others depriving you of your liberty-right to practice whichever religion you choose.

11. What is the relationship between rights and right action?

(a) The consequentialist answer: the violations of rights is a bad thing. It should be folded into the consequences of an action; and the action is right iff it maximizes the goodness of the consequences.

i. This means that it can be right to kill one person to prevent two from being killed. For, in this case, you face a choice between one person's right to life being violated and two people's rights to life being violated. Assuming that two rights violations are worse than one comparable right violation, it will be best to kill the one.

(b) The Nozickian answer (a non-consequentialist answer): rights provide *absolute side-constraints* on right action. In order for an act to be right, it must not violate anybody's rights. Even if killing one to save two would minimize the total number of rights violations, it is *still wrong* to violate somebody's right to life.

8.2.3 Justice as Respecting Rights

12. For the libertarian, justice is entirely a matter of respecting rights. A just state cannot infringe upon the rights of its citizens. A just state must similarly provide means to protect the rights of its citizens—*e.g.*, a criminal justice system.

(a) Much of the content of the view depends upon the particular rights it believes individuals have.

(b) It is possible to be a libertarian and to think, *e.g.*, that individuals have so-called *positive* rights to health care, to sustenance, and so on.

(c) One prominent kind of right which typical libertarians stress is the right to the ownership of property and the economic right to voluntarily transfer ownership.

i. Locke held that we all hold natural rights over our own bodies and our own labor. By "mixing our labor" with an unowned resource, we can thereby come to own that resource as well. For instance, if a farmer works an unowned field, they thereby can come to own the field. If a carpenter cuts down an unowned tree and builds a chair, they thereby come to own that chair.
   A. Locke believed that you had to leave a good amount of the land, trees, *etc*. behind for others to use when you do this.

ii. Nozick holds that we can come to justly own some property if:
   A. it is previously unowned
   B. by coming to acquire it, we do not violate anybody's rights; and
   C. by coming to acquire it, we do not leave anybody else worse off.

iii. Nozick holds that we can justly transfer property (e.g., by trade) iff the transfer does not involve any rights violations.

8.3 Egalitarianism

What is Egalitarianism? What is the 'levelling down' objection to egalitarianism? What is the view that Anderson calls 'Luck Egalitarianism'? Describe one of her arguments against it. What is Anderson's alternative version of Egalitarianism, and how does it address the objection to Luck Egalitarianism you discussed?

1. An *Egalitarian* is anyone who holds that some kinds of inequalities are unjust (and, therefore, society has a duty to correct certain kinds of inequalities).

   (a) Different versions of egalitarianism will disagree about which kinds of inequalities are unjust, and why they are unjust.

2. The first question for the Egalitarian: what kinds of inequalities are unjust? (What do we have an obligation to equalize?)

   (a) First answer: inequalities in *welfare* are unjust (*ie*, we have an obligation to equalize welfare).
      i. An objection: Assume preferentism about welfare. Then, some people have expensive tastes, and the only way to equalize welfare would be to direct additional resources to these people. (*Cf*. Hausman's objection to preferentism.)

   (b) Second answer: inequalities in *resources* are unjust (*ie*, we have an obligation to equalize resources).
      i. A clarificatory question: what do we mean by an equal distribution of resources? We presumably shouldn't want an equal distribution of winter coats. Winter coats are needed more in Canada than they are in Brazil.
      ii. An answer: we should begin by providing everyone with an equal share of all resources, and then allow trades. We will thereby arrive at an 'envy-free' distribution.

3. The second question for Egalitarianism: *why* are inequalities unjust?

   (a) We could think that certain inequalities *per se* are not unjust, but rather than those inequalities are *evidence for* injustice, or perhaps that they *cause* injustice.

(b) Alternatively, we could think that certain inequalities are *intrinsically* unjust.

(c) A *first pass* objection to egalitarianism (the '*leveling down*' objection): if inequality (in either welfare or resources) is intrinsically unjust, then it could be that justice demands that we 'level down' by reducing the resources/welfare of those at the top without increasing the resources/welfare of those at the bottom. But justice does not demand that we 'level down'. So inequality is not intrinsically unjust.

8.3.1 Luck Egalitarianism

4. Another *first pass* objection to egalitarianism: some disparities in welfare are the result of irresponsibility. Suppose, for instance, that somebody squanders all their money at the Casino. Society does not owe this person compensation to correct their self-imposed misfortune. If egalitarianism commits us to saying that this person is owed compensation for their losses at the casino, then egalitarianism is false.

   (a) A reply: properly understood, egalitarianism doesn't say that *all* inequalities in welfare/resources are unjust. Rather, it just says that *undeserved* inequalities are unjust.

   (b) Following Anderson, call this brand of egalitarianism *Luck* Egalitarianism.

   > Luck Egalitarianism
   > Undeserved inequalities—that is, inequalities arising from brute luck alone—are unjust. Therefore, society has an obligation to compensate the victims of undeserved bad luck with funds taken from the beneficiaries of undeserved good luck.

5. To understand Luck Egalitarianism, we should first understand its distinction between *brute luck* and *option luck*.

   (a) *Brute luck* is luck which determines people's initial endowments of resources or opportunities—it is luck over which the individual has no control. Brute luck accounts for my androgenic alopecia, e.g. It also accounts for the education I received, the amount of money my parents had, how well fed I was as a child, and so on.

   (b) *Option luck* is luck which results from individual's free choices—luck over which the individual had some control. For instance, the person who loses all their income at the Casino has had bad *option* luck, not bad *brute* luck. Similarly, the person who is injured while rock climbing suffers from bad *option* luck, not bad *brute* luck.

   (c) Where insurance markets exist, they transform some kinds of brute luck into option luck. If you fail to buy into the health insurance market, *e.g.*, and then get sick but lack insurance, then you face bad luck which is the result of your failure to purchase health insurance—*i.e.*, bad option luck. (While, if there had been no market for health insurance, then your sickness would have been an instance of bad brute luck.)

6. The Luck Egalitarian thinks that society should provide a kind of insurance market for bad brute luck.

7. Anderson objects to luck egalitarianism on two grounds. Firstly, she thinks that luck egalitarianism expresses disrespect and pity for the victims of bad brute luck.

(a) If luck egalitarianism is correct, then the disabled, for instance, are in a position to make claims on others to provide them with accommodation in virtue of their *inferiority*.

(b) Anderson thinks that, rather, the disabled are in a position to make claims on others to provide them with accommodation in virtue of their *equality*.

8. Secondly, Anderson thinks that the luck egalitarian delivers the wrong verdicts about victims of bad option luck.

   (a) Consider cases like the following:

   > Uninsured Driver
   > An uninsured driver, Sally makes an illegal turn and gets into an accident.

   (b) Anderson believes that justice demands that we provide Sally with medical care. However, the luck egalitarian has nothing to say about why Sally is owed medical care.

   > Negligent Disability
   > After the accident, the Sally is severely disabled.

   (c) Anderson believes that we owe assistance to Sally—*e.g.*, by providing access ramps for her wheel chair. However, according to the luck egalitarian, we owe Sally nothing, because her misfortune is the result of bad option luck.

   > Dangerous Job
   > Bob works a dangerous job and gets injured.

   (d) Anderson believes that Bob, just like Sally, is owed medical care and disability accommodation. However, since their labor was voluntarily provided, their injuries are a case of bad option luck, and therefore the luck egalitarian says that we owe them nothing.

8.3.2 Democratic Egalitarianism

9. In place of luck egalitarianism, Anderson endorses *Democratic Egalitarianism*.

   (a) In response to the question 'which inequalities are unjust?', the democratic egalitarian says that inequalities in *social status* are unjust. What we owe to each other is *equal treatment*—to treat all members of society as equals; but, more than just equal treatment, we owe everybody the resources necessary for them to develop the capabilities necessary for them to genuinely function as democratic equals.

   > Democratic Egalitarianism
   > Inequalities in social status are unjust. Therefore, society has an obligation to both (a) treat everyone as an equal; and (b) provide people with the resources necessary for them to have the *opportunity* to develop the *capacity* to function as an equal citizen.

      i. Part of what is involved in treating somebody as a democratic equal is thinking that our justification for our decisions must be acceptable to them.

   (b) In response to the question 'why are these inequalities unjust?', the democratic egalitarian says that we owe each other equal treatment (and the opportunity to achieve equal standing) for the very same reason that we own everyone an equal vote: this is one of the basic preconditions of a democratic society.

10. The democratic egalitarian will not say that inequalities in distribution of resources or welfare are *intrinsically* unjust. However, inequalities in distribution of resources or welfare naturally lead to social hierarchies, and so the democratic egalitarian has *instrumental* reason to oppose inequalities in resources and welfare.

11. Consider what the democratic egalitarian says about the following cases:

    Uninsured Driver
    An uninsured driver, Sally makes an illegal turn and gets into an accident.

    (a) We owe Sally medical care because without such care, Sally will not be capable of standing as a free and equal member of society. This is owed to Sally irrespective of whether the accident was the result of her own poor choices.

    Negligent Disability
    After the accident, the Sally is severely disabled.

    (b) We similarly owe Sally wheel chair accessible ramps (*e.g.*), because without these, she will be denied access to public spaces and be incapable of participating as an equal member of society.

Chapter 9

Introduction to Social Welfare Functions

1. What's happened so far: the framework of Welfare Economics has built up an account of *individual* welfare. That is given by actual preferentism:

   > Actual Preferentism
   > Each individual, i, has a *preference ordering* over outcomes, $\succeq_i$, and i's welfare is entirely a function of how well satisfied i's preferences are.
   >
   > (a) That is, if $A \succeq_i B$, then the outcome A is at least as good for i than the outcome B is; and
   >
   > (b) if $A \succ_i B$, then the outcome A is better for i than the outcome B is.

2. We additionally assumed that the *goodness* of an outcome is determined entirely by the levels of welfare of every individual in the society:

   > Welfarism
   > The *goodness* of an outcome is entirely a function of the welfare of the individuals in that outcome; and higher levels of welfare are better.

3. Actual Preferentism and Welfarism entail what we've called the Pareto Principle,

   > Pareto Principle
   > For any outcomes A, B, if some prefer A to B and none prefer B to A, then A is *better* than B.

4. Here's how we should think about this principle: we've got some *individual* preference orderings, $\succeq_i$. From this, we want to cook up a *group* preference ordering—or a *social betterness* ordering, $\succeq_G$. This social betterness ordering will, given Welfarism, tell us something about the level of welfare of society as a whole.

   (a) That is, if $A \succeq_G B$, then outcome A is at least as good as outcome B (not just better for any particular individual, but better *full stop*).

   (b) We can define $A \succ_G B \stackrel{\text{def}}{=} A \succeq_G B \;\&\; B \not\succeq_G A$, and similarly define $A \sim_G B \stackrel{\text{def}}{=} A \succeq_G B \;\&\; B \succeq_G A$.

   (c) If $A \succ_G B$, then outcome A is better than outcome B (not just better for any particular individual, but better *full stop*).

(d) If $A \sim_G B$, then outcome A is *just as good* as the outcome B (not just as good for any particular individual, but just as good *full stop*).

5. Call any function from individual preferences to group preferences a SOCIAL WELFARE FUNCTION.

   SOCIAL WELFARE FUNCTION
   A SOCIAL WELFARE FUNCTION W is a function from *individual preference orderings* to *group preference orderings*.

$$W : \begin{bmatrix} \succeq_1 \\ \succeq_2 \\ \vdots \\ \succeq_i \\ \vdots \\ \succeq_n \end{bmatrix} \quad \rightarrow \quad \succeq_G$$

6. The Pareto Principle only gives us a *partial* group ordering over outcomes.[1] That is: the ordering it gives us over outcomes only relates *some* outcomes; it does not relate *all* outcomes.

   (a) We tried to further enrich this social preference relation by using the notion of Kaldor-Hicks efficiency, but we saw that this preference relation didn't meet the minimal constraints of being reflexive and transitive.

7. Here's an idea: let's think about the individuals in society as each having a *vote* on what the outcome will be.

   (a) Assume that individuals will vote their conscience—that is, they will reveal their actual preferences in their voting behavior.

   (b) Then, define a *social or group betterness ordering*, $\succeq_G$, using a *voting rule*.

   (c) That is: our *social welfare function* can just be a *voting rule*.

9.1 Voting Rules

1. Suppose that everybody hands us their preference ordering over all of the outcomes. We may compile this information by keeping track of how many of each preference ordering we receive. For instance, if 7 people prefer A to B to C, 6 people prefer B to A to C, 5 people prefer C to A to B, and 4 people prefer C to B to A, then we may represent this with the following table, which we can call a *voter profile*:

	# of Votes			
	7	6	5	4
1st	A	B	C	C
2nd	B	A	A	B
3rd	C	C	B	A

   This voter profile tells us everything about the preferences of all of the individuals in the society.

[1] Strictly speaking, this is a *pre-order*, and not a partial order. That is, the order is reflexive and transitive, but not total.

9.1.1 Plurality Method

2. Given this information, how do we decide upon which option to choose? By far and away the most prevalent answer to this question (amongst non-experts) is given by the PLURALITY METHOD.

   PLURALITY METHOD
   The option which is in the 1st position of *most* people's preference ordering is best. The option which has the second-most number of 1st position votes is second best. The option which has the third-most number of 1st position votes is third best; and so on.

   Notice that the plurality method does not care about anything other than people's first choice.

   (a) In the example above, the group preference ordering would be given as follows:

   $$C \succ_G A \succ_G B$$

   since C got 9 votes, A got 7 votes, and B got 6 votes.

3. Suppose that all voters who preferred A to B change their mind and begin to prefer B to A. But nobody changes their preference between B and C (nor between A and C). Then, we would have the following voter profile:

	# of Votes	
	13	9
1st	B	C
2nd	A	B
3rd	C	A

   Then, the plurality method would give us the following group preference ordering:

   $$B \succ_G C \succ_G A$$

   since B gets 13 votes, C gets 9 votes, and A gets 0 votes.

   Notice what just happened: everybody who *previously* preferred B to C *still* prefers B to C. And everybody who *previously* preferred C to B *still* prefers C to B. All that changed was how people rank A and B relative to each other. But, while the plurality method previously said that the group prefers C to B, the plurality method *now* says that the group prefers B to C.

   (a) What we've just learned is that, if we form *group* preferences out of individual preferences by using the plurality method, then our group preferences will not satisfy the principle known as INDEPENDENCE OF IRRELEVANT ALTERNATIVES.

   INDEPENDENCE OF IRRELEVANT ALTERNATIVES
   Whether the group prefers X to Y is determined entirely by each individual's relative ranking of X and Y.

   The PLURALITY METHOD violations this principle since changing how voters rank the irrelevant option A changes what the plurality method says about the relative ranking of B and C.

4. We've seen that, given this voter profile,

	# of Votes			
	7	6	5	4
1st	A	B	C	C
2nd	B	A	A	B
3rd	C	C	B	A

C would win in a plurality election between A, B, and C. Note, though, that A would win in a one-on-one contest between just A and C, for, in that case, we'd have the following voter profile:

	# of Votes	
	13	9
1st	A	C
2nd	C	A

What's more, A would win in a plurality election between A and B. Since, if we asked people to just rank their preference between A and B, we'd get:

	# of Votes	
	12	10
1st	A	B
2nd	B	A

(a) What this tells us is that A is the Condorcet winner. A Condorcet winner is an option which wins in a one-on-one contest with every other option on the menu.

> Condorcet Winner
> Given a menu of options, X is the Condorcet winner if, *for every other option on the menu*, given a choice between X and that option, most people prefer X.

(b) What we've learned is that a plurality voting method can fail to select the Condorcet winner. For, in the above example, A is the Condorcet winner, yet A loses to C in an election between A, B, and C.

(c) One plausible constraint on a voting method is the *Condorcet Winner Criterion*:

> Condorcet Winner Criterion
> If there is a Condorcet winner, then they must be first in the group's preference ordering.

The Plurality method fails the Condorcet Winner Criterion, as we've just seen. A is the Condorcet winner, yet they are not first in the social group preference ordering determined by the Plurality method.

5. Notice something else: C would lose to A in a one-on-one race. Similarly, C would lose to B in a one-on-one race. For, if A is removed from the menu, then we'd have the following voter profile:

	# of Votes	
	13	9
1st	B	C
2nd	C	B

So the plurality method tells us that $B \succ_G C$.

(a) That means that C is a *Condorcet loser*. They lose against *every other candidate* in a one-on-one race. What we've learned is that Plurality methods can select Condorcet losers.

> Condorcet Loser
> Given a menu of options, X is the Condorcet Loser iff, *for every other option on the menu*, given a choice between X and that option, most people prefer the other option.
>
> Condorcet Loser Criterion
> If there is a Condorcet Loser, then they cannot be selected as the group's top preference.

(b) Plurality methods fail to satisfy the Condorcet Loser Criterion.

6. To summarize what we've learned thus far:

	Voting Method
CRITERION	Plurality
Independence of Irrelevant Alternatives	×
Condorcet Winner Criterion	×
Condorcet Loser Criterion	×

9.1.2 The Condorcet Paradox

7. Why not just select the Condorcet winner? This would guarantee that we satisfy the Condorcet winner criterion. One reason is that there is not always guaranteed to exist a Condorcet winner. For instance, consider the following voter profile.

	# of Votes		
	1	1	1
1st	A	C	B
2nd	B	A	C
3rd	C	B	A

(a) Suppose that there were a race between A and B. Then, A would win 2 to 1.

(b) Suppose that there were a race between A and C. Then, C would win 2 to 1.

(c) Suppose that there were a race between B and C. Then, B would win 2 to 1.

8. In this election, every option loses to some other option in a one-on-one race. So there is no Condorcet winner (nor a Condorcet loser).

9. Not only that—but, if we assume that, if X is preferred to Y by 2/3rds of voters, then the group prefers X to Y, then we get the following absurd result: the group prefers A to B, and B to C, and C to A.

$$A \succ_G B \succ_G C \succ_G A$$

So: this seemingly reasonable assumption leads to the absurd consequence that group preference is *cyclic*. This is known as 'Condorcet's Paradox'.

9.1.3 Instant Runoff Method

10. Here is a way of attempting to fix up the problems with the plurality method. In the example we considered previously, C was selected as the group's top preference even though they were a Condorcet loser. Intuitively, the reason for this was that the anti-C-vote was *split* between those who preferred A and those who preferred B.

11. A potential solution is given by the INSTANCE RUN-OFF METHOD of voting.

    > INSTANT RUN-OFF METHOD
    > We calculate how many first-place votes each option receives. The option with the least first-place votes is eliminated from the menu of options. Then, with the restricted menu of options, we count how many first-place votes each option gets. The option with the least first-place votes in this second round of voting is eliminated. We proceed in this manner until one option has a majority. This is the top of the group preference ordering. To determine the ranking of other options, remove the top preference and proceed as before.[2]

12. Let's apply the Instant Run-Off Method to our previous case.

	# of Votes			
	7	6	5	4
1st	A	B	C	C
2nd	B	A	A	B
3rd	C	C	B	A

    (a) In round 1, we look at people's first choice between A, B, and C. A gets 7 votes, B gets 6, and C gets 9. So B is eliminated.

    (b) In round 2, we look at people's first choice between A and C. A gets 13 votes and C gets 9. So A has a majority, and A wins.

    (c) Between B and C, most prefer B. So, our group preference ordering will then be:

$$A \succ_G B \succ_G C$$

13. The Instant Run-Off method is helpful for mitigating vote-splitting. However, it will not *always* satisfy the independence of irrelevant alternatives. Let's imagine that the voters in column 3 changed their minds and decided that, actually B is preferable to A. Then, we'd have the following voter profile:

	# of Votes		
	7	6	9
1st	A	B	C
2nd	B	A	B
3rd	C	C	A

    (a) Now, B is the Condorcet winner.

[2] Note that this procedure says nothing about what to do if there is a tie for last place. We would have to include such a specification to get a voting method that returns a definite verdict in every possible case.

i. In a choice between A and B, B wins 15 to 7.

ii. In a choice between B and C, B wins 13 to 9.

(b) However, the winner of an instant run-off method would *still* be A. Things would proceed exactly as before:

i. In round 1, we look at people's first choice between A, B, and C. A gets 7 votes, B gets 6, and C gets 9. So B is eliminated.

ii. In round 2, we look at people's first choice between A and C. A gets 13 votes, and C gets 9. So A has a majority, and A wins.

14. So the Instant Run-Off method fails to satisfy the Condorcet Winner Criterion.

15. Instant Run-Off Voting does, however, satisfy the Condorcet Loser Criterion. (Can you see why?)

16. Here is a very odd consequence of Instant Run-Off methods:

(a) Suppose that we begin with the following voter profile:

	# of Votes		
	7	6	7
1st	A	B	C
2nd	B	A	B
3rd	C	C	A

i. In round 1, A and C both get 7 votes, while B gets 6. B is eliminated.

ii. In round 2, A gets 13 votes, and C gets 7. A is selected.

(b) Suppose that two of the C voters changes their minds, and decides that A is actually their top preference, then C, and then B. We will then have the following voter profile:

	# of Votes			
	7	6	5	2
1st	A	B	C	A
2nd	B	A	B	C
3rd	C	C	A	B

(c) A was previously the top group preference, and all that's changed is that A has moved up in some voter's preference ranking. But *now*, B is the group's top preference.

i. In round 1, A gets 9 votes, B gets 6, and C gets 5. So C is eliminated.

ii. In round 2, A gets 9 votes and B gets 11. So B has a majority, and B is selected.

17. What we've learned is that Instant Run-Off methods fail a criterion called 'Monotonicity'.

> Monotonicity
> It should not be possible to lower an option's ranking in the group order by raising it in some individual's ranking and otherwise leaving all preferences unchanged.

(a) In contrast, the Plurality method will always satisfy Monotonicity (Can you see why?)

18. A summary:

	Voting Method	
CRITERION	Plurality	Instant Run-Off
Independence of Irrelevant Alternatives	×	×
Condorcet Winner Criterion	×	×
Condorcet Loser Criterion	×	✓
Monotonicity	✓	×

9.1.4 Copeland's Method

19. Here's a method which is guaranteed to satisfy both the Condorcet Winner Criterion and the Condorcet Loser Criterion, as well as the Monotonicity criterion: you look at the number of victories and losses each option would win in every one-on-one race. Options are then ordered according to their total number of wins minus their total number of losses.

    COPELAND'S METHOD
    Consider every one-on-one contest between each pair of options. For each option, count the number of its wins and the number of its losses. Assign it a score which is the sum of its wins, minus its losses. The group's preference ordering is given by the order of each option's score

20. For instance, consider the following voter profile:

	# of Votes			
	7	6	4	8
1st	A	B	D	A
2nd	B	A	B	C
3rd	C	D	A	D
4th	D	C	C	B

    (a) There are 6 possible one-on-one races, and the outcomes are as follows:

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} - & A & A & A \\ A & - & B & B \\ A & B & - & C \\ A & B & C & - \end{bmatrix} \end{array}$$

    i. A wins three one-on-one contests and loses none; so A's score is 3 - 0 = 3
    ii. B wins two one-on-one contest and loses one; so B's score is 2 - 1 = 1
    iii. C wins one one-on-one contest and loses twp; so C's score is 1 - 2 = -1
    iv. D wins no one-on-one contest and loses three; so D's score is 0 - 3= -3
    v. Thus, the group's preference ordering is given by

$$A \succ_G B \succ_G C \succ_G D$$

    (b) A is the Condorcet winner, and A is at the top of the group preference ranking.

    (c) D is the Condorcet loser, and D is at the bottom of the group preference ranking.

(d) Moreover, this isn't an accident: Copeland's method will *always* have the Condorcet winner at the top of the group ranking, if it exists. Moreover, all Condorcet losers will be at the bottom of the group ranking. (Can you see why?)

21. We won't prove it here, but it also turns out that Copeland's method satisfies MONOTONICITY. Let's check this by looking at the case we considered earlier. There, the original voting profile was

	# of Votes		
	7	6	7
1st	A	B	C
2nd	B	A	B
3rd	C	C	A

(a) There are three possible one-on-one contests, and the outcomes are as follows:

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} - & B & A \\ B & - & B \\ A & B & - \end{bmatrix} \end{array}$$

i. B's score is $2 - 0 = 2$
ii. A's score is $1 - 1 = 0$
iii. Cs score is $0 - 2 = -2$
iv. So, given the Copeland method, the group preference ordering is:

$$B \succ_G A \succ_G C$$

(b) If we then have 3 of the C-voters change their mind and decide that A is actually their top preference, then we will have the following voter profile:

	# of Votes			
	7	6	4	3
1st	A	B	C	A
2nd	B	A	B	C
3rd	C	C	A	B

(a) Now, there are again three possible one-on-one contests, and the outcomes are:

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} - & \text{tie} & A \\ \text{tie} & - & B \\ A & B & - \end{bmatrix} \end{array}$$

i. B's score is $1 - 0 = 1$
ii. A's score is $1 - 0 = 1$
iii. Cs score is $0 - 2 = -2$
iv. So, given the Copeland method, the group preference ordering is:

$$B \sim_G A \succ_G C$$

(b) So the Copeland method does not, in this case at least, lead to a violation of Monotonicity in the way that Instant Run-off Voting did.

(c) And this is no accident. In fact, in general, Copeland's method will always satisfy Monotonicity.

22. What about the independence of Irrelevant Alternatives? Consider the voting profile from 'Condorcet's Paradox' again:

	# of Votes		
	1	1	1
1st	A	C	B
2nd	B	A	C
3rd	C	B	A

(a) There are three possible one-on-one contests, and the outcomes are as follows:

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} - & A & C \\ A & - & B \\ C & B & - \end{bmatrix} \end{array}$$

i. A wins one and loses one, so A's score is $1 - 1 = 0$

ii. B wins one and loses one, so B's score is $1 - 1 = 0$

iii. C wins one and loses one, so C's score is $1 - 1 = 0$

iv. Thus, the group's preference ordering is given by

$$A \sim_G B \sim_G C$$

(b) However, suppose that everyone who previously preferred A to B changes their mind, and ranks B above A—but they don't change their ranking of A and C (or of B and C). Then, we'd have the following voter profile:

	# of Votes		
	1	1	1
1st	B	C	B
2nd	A	B	C
3rd	C	A	A

i. Then, there are three possible one-on-one contests, and the outcomes are as follows:

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} - & B & C \\ B & - & B \\ C & B & - \end{bmatrix} \end{array}$$

A. B wins two and loses none, so B's score is $2 - 0 = 2$

B. C wins one and loses one, so C's score is $1 - 1 = 0$

C. A wins none and loses two, so A's score is $0 - 2 = -2$

D. Thus, the group's preference ordering is given by

$$B \sim_G C \sim_G A$$

(c) Even though nobody changed their relative ranking of A and C, the group went from being *indifferent* between A and C to preferring C to A.

(d) So the Copeland method violates the Independence of Irrelevant Alternatives.

23. A summary:

	Voting Method		
CRITERION	Plurality	Instant Run-Off	Copeland
Independence of Irrelevant Alternatives	×	×	×
Condorcet Winner Criterion	×	×	✓
Condorcet Loser Criterion	×	✓	✓
Monotonicity	✓	×	✓

9.1.5 Caveats

24. There are tons of other voting methods out there, and tons of other criteria that people think voting systems should satisfy. We have probably not discussed the best voting procedure, or the most important criteria. This has been no more than a brief introduction to an incredibly complex field.

25. We should also not think that our choice between voting systems is accomplished by counting up the number of criteria they satisfy. We should also think about, for instance, how likely the system is to yield violations of the criteria, and several other considerations, including, e.g., whether the system is susceptible to easily implemented strategic voting (as plurality voting and instant run-off voting plainly are).

9.2 Arrow's Impossibility Result

26. We've laid down some criteria that it seems reasonable (*prima facie*) to want a voting system to have. Can we have them all? There's a famous result by Kenneth Arrow which shows that we *cannot*.

> ARROW'S IMPOSSIBILITY THEOREM
> If there are 3 or more options, there is no social welfare function from individual voter preferences to group preference orderings which satisfies all of the following criteria:
>
> > UNANIMITY
> > If every individual has exactly the same preference ordering, then that is the group's ordering.
> >
> > MONOTONICITY
> > It is not possible to lower an option's ranking in the group preference ordering by raising it in some individual's ranking and otherwise leaving all preferences unchanged.

Independence of Irrelevant Alternatives
Whether the group prefers X to Y is determined entirely by each individual's relative ranking of X and Y.

No Dictator
The social welfare function should not be *dictatorial*—that is, it should not make one individual's preference ordering the group preference ordering no matter what other's preferences happen to be.

27. In the presence of Independence of Irrelevant Alternatives, Unanimity and Monotonicity are equivalent to Weak Pareto. This affords another version of the theorem:

Arrow's Impossibility Theorem (ver 2)
If there are 3 or more options, there is no social welfare function from individual voter preferences to group preference orderings which satisfies all of the following criteria:

Weak Pareto
If everyone prefers X to Y, then the group prefers X to Y.

Independence of Irrelevant Alternatives
Whether the group prefers X to Y is determined entirely by each individual's relative ranking of X and Y.

No Dictator
The social welfare function should not be *dictatorial*—that is, it should not make one individual's preference ordering the group preference ordering no matter what other's preferences happen to be.

Chapter 10

Social Welfare Functions: *Liberté et Égalité*

1. Recall, the welfare economist accepts *preferentism* and *welfarism*. This means that they believe the goodness of a state of affairs is determined by everyone's individual preferences. Their views about how this happens can be represented with a *social welfare function*, G, from individual's preference orderings to a group preference ordering.

   Social Welfare Function
   A social welfare function G is a function from *individual preference orderings* to *group preference orderings*.

$$G : \begin{bmatrix} \succeq_1 \\ \succeq_2 \\ \vdots \\ \succeq_i \\ \vdots \\ \succeq_N \end{bmatrix} \quad \rightarrow \quad \succeq_G$$

   (a) As a bare minimum, the preferentist welfarist will want a social welfare function which satisfies the **Weak Pareto** principle.

   Weak Pareto If everyone prefers S to T, then S is better than T.

$$\text{if, for all } i,\ S \succ_i T,\ \text{then } S \succ_G T$$

10.1 *Liberté*

What is Sen's argument against the possibility of a Paretian Liberal? Clearly state the three principles which Sen shows to be incompatible, and illustrate the conflict by explaining the example of Prude and Lewd. Then, introduce Allan Gibbard's example (called 'Match and Clash' in class), and explain why it shows that Sen's conflict can be generated without assuming the Pareto Principle.

2. Recall, the *libertarian* thought that justice was entirely a matter of respecting people's liberties. In particular, the state should not infringe upon people's freedom to own and freely trade property—that is, it should not interfere with the *free market*.

   (a) We also saw that the free market will always take us to a *Pareto optimum*.

3. Can we incorporate a commitment to *liberty* into our social welfare functions?

4. As a first pass, let's suppose that we want our social welfare function to leave at least *some* options up to the discretion of the individual.

   (a) When it comes to certain matters—like, *e.g.*, whether Dmitri has an ear piercing—we want our social welfare function to say that, if Dmitri prefers to have his ears pierced, then it is better that he have his ears pierced. And if Dmitri prefers to have his ears unpierced, then it is better that he not have his ears pierced.

   (b) That is, we're going to suppose that our social welfare function honors the individual i's freedom to decide between the options S or T just in case, when it comes to the choice between S and T, individual i's preference *becomes* the group preference.

   (c) Given a social welfare function, let's say that the individual i's preference is *decisive* in the choice between S and T iff:

      i. if $S \succ_i T$, then $S \succ_G T$;
      ii. if $S \sim_i T$, then $S \sim_G T$; and
      iii. if $T \succ_i S$, then $T \succ_G S$.

   (d) As a first pass, then, let's say that we want a social welfare function to honor the principle of **Minimal Liberalism**:

      Minimal Liberalism There are at least two individuals, i and j, and at least two pairs of options, (S_i, T_i) and (S_j, T_j), such that individual i is decisive in the choice between S_i and T_i, and individual j is decisive in the choice between S_j and T_j.

      i. This is an incredibly minimal form of liberalism. A social welfare function meeting this constraint doesn't honor *everybody*'s freedom. Instead, it only honors *two* people's freedom. And it only honors their freedom two choose between *two* outcomes. So it looks like anyone who values freedom should accept **Minimal Liberalism**.

         A. A quick aside: I am also, with the welfare economist, assuming *consequentialism*, so I am assuming that the reason the liberal thinks it is wrong to deprive people of liberty is that doing so makes them *worse off*. You could be a libertarian, like Nozick, who rejects consequentialism. In that case, your libertarianism need not commit you to **Minimal Liberalism**.

5. We'll also suppose that, in addition to **Minimal Liberalism**, we want our social welfare function to satisfy **Weak Pareto**.

6. And, of course, we want the group preference ordering to have some basic structural properties—we don't want it to deliver cycles of betterness, for instance.

   No Cycles There is no sequence of states-of-affairs, $S_1, S_2, S_3, \ldots, S_N$ such that S_1 is better than S_2, which is better than S_3, which is better than $\ldots$, which is better than S_N, which is better than S_1.

$$S_1 \succ_G S_2 \succ_G S_3 \succ_G \cdots \succ_G S_N \succ_G S_1$$

10.1.1 The Impossibility of a Paretian Liberal

7. These three principles each look very plausible at a first pass. However, Amartya Sen showed that we cannot accept all three at once.

   Sen's Theorem
   There is no social welfare function which satisfies **Minimal Liberalism**, **Weak Pareto**, and **No Cycles**

8. We won't go through the formal proof of this theorem, but we can look at an illustrative example which Sen provides (I've updated the cultural references):

   Prude and Lewd
   Prude is outraged and offended by *Fifty Shades of Gray*. Lewd, on the other hand, is delighted by the book. Prude would most prefer that nobody read the filth. However, if somebody must read it, Prude would rather read it himself than expose a libertine like Lewd to its influence. Lewd would most prefer that both he and Prude read the book. However, if only one of them is to read it, Lewd would rather it be Prude—he relishes the thought of Prude's horrified reactions.
   Thus, Prude and Lewd's preference orderings are given in the following table.

	Prude	Lewd
1st	Neither reads (N)	Both read (B)
2nd	Prude reads (P)	Prude reads (P)
3rd	Lewd reads (L)	Lewd reads (L)
4th	Both read (B)	Neither reads (N)

   (a) For the purposes of illustration, suppose that Prude and Lewd are the only people in this society. And let's suppose that a liberal thinks that whether to read a book is a matter which ought to be left up to the individual. If you prefer to read, then it's better that you read; if you prefer to not read, then it's better that you refrain. That's all we need to bring out the conflict between **Weak Pareto**, liberalism, and **No Cycles**.

   (b) The only difference between P and N is whether Prude reads. It should be entirely up to Prude whether he read or not. Since he prefers to not read, the liberal says that it is better if he doesn't. So

   $$N \succ_G P$$

   (c) Note that both Prude and Lewd prefer P to L. So, by **Weak Pareto**,

   $$P \succ_G L$$

   (d) And, the only difference between L and N is whether Lewd reads. It should be entirely up to Lewd whether he read or not. Since he prefers to read, the liberal tells us that it is better if he does. So

   $$L \succ_G N$$

   (e) But now we've violated **No Cycles**.

   $$N \succ_G P \succ_G L \succ_G N$$

9. So it seems that we face a choice: either reject the **Weak Pareto**, or reject **Minimal Liberalism**. You can't be both a Paretian and a liberal. If you want to be a liberal, you'd better reject the Weak Pareto principle. If you want to be a Paretian, you'd better reject **Minimal Liberalism**.

10.1.2 The Impossibility of a Liberal?

10. Actually, matters are worse. Even rejecting **Weak Pareto Principle** won't get you out of trouble. Allan Gibbard showed that liberalism leads to cycles all by itself, even without **Weak Pareto Principle**.

11. Consider the following case, from Gibbard:

> Match and Clash
> Clash is a non-conformist. She would prefer having a pierced nose, but what's most important to her is that her fashion be different from Match's. So she wants to pierce her nose if (but only if) Match doesn't pierce hers. Match is a follower. She doesn't want to pierce her nose, but she does want her fashion to match Clash's. So she wants to pierce her nose if (and only if) Clash pierces hers.
> Therefore, Clash and Match's preferences are given by the following table.

	Clash	Match
1st	Clash pierces (C)	Neither pierce (N)
2nd	Match pierces (M)	Both pierce (B)
3rd	Both pierce (B)	Clash pierces (C)
4th	Neither pierce (N)	Match pierces (M)

(a) For the purposes of illustration, let's suppose that Match and Clash are the only people in this society. And let's suppose that whether your nose is pierced is the kind of thing which ought to be left up to the individual. If you prefer a pierced nose, then it's better if your nose is pierced. And if you don't, then it's better if it's not pierced.

(b) Note that the only difference between C and N is whether Clash pierces. And Clash prefers N to C. So the liberal says that C is better than N.

$$C \succ_G N$$

(c) The only difference between N and M is whether Match pierces. And Match prefers N to M. So the liberal says that N is better than M.

$$N \succ_G M$$

(d) The only difference between M and B is whether Clash pierces. And Clash prefers M to B. So the liberal says that M is better than B.

$$M \succ_G B$$

(e) And, finally, the only difference between B and C is whether Match pierces. Since Match prefers B to C, the liberal says that B is better than C.

$$B \succ_G C$$

(f) But now we've contradicted **No Cycles**

$$C \succ_G N \succ_G M \succ_G B \succ_G C$$

(g) So the kind of liberalism represented by **Minimal Liberalism** is inconsistent with **No Cycles** all by itself. We didn't have to bring up the **Weak Pareto** princple at all.

10.1.3 The Impossibility of a Paretian Liberal, take 2

12. Sen's principle **Minimal Liberalism** assumes that the right way to think about liberalism is in terms of the decisiveness of individual preference. Sen's liberal thinks that, if Prude prefers to not read, then it's better if Prude not read (all else equal). And, if Lewd prefers to read, then it's better if Lewd read (again, all else equal). What Gibbard's case shows us, I think, is that this way of understanding liberalism is misguided.

    (a) And, in retrospect, we should recognize that we ought to have rejected **Minimal Liberalism** as a characterization of liberalism on independent grounds. Liberals think that certain self-regarding decisions should be left up to the individual. One of Mill's arguments for this was that the individual is in a better position to know what's best for them than the rest of society. However, Mill didn't think that individuals were necessarily *right* about what's in their own best interest.

    (b) Liberals should acknowledge that people can and do make self-regarding choices that make them worse off. A libertine liberal like Lewd will grant that what Prude reads should be left up to him. But that won't keep Lewd from thinking that Prude's diet of religious drivel is making his life worse. And surely Lewd should also think that, all else equal, it's better for people's lives to go better, and worse for them to go worse.

    (c) So the liberal shouldn't think that, when it comes to self-regarding decisions, people's actual preferences are objectively best. What they should think is that, when it comes to self-regarding decisions, it is better to allow people to choose for themselves than for the state to choose for them. That is: the liberal ought to think that, when it comes to self-regarding decisions, it is worse to deprive people of liberty than it is to allow them to make their own poor choices.

        i. At least, this is what a consequentialist liberal ought to think—though a non-consequentialist liberal is free to admit that things would be better if people were compelled to make the right choices, though such an arrangement would be unjust in spite of its betterness.

13. Our earlier discussion did not draw any distinction between possibilities in which people were forced to take certain options and those in which they freely chose those options. Let's introduce this distinction, and use it to formulate a new principle of liberalism.

    Liberalism as Non-Compulsion For any two states F and U, if, in F, all self-regarding choices were *free* and, in U, some self-regarding choice was *unfree*, then $F \succ_G U$.

    (a) According to this principle, when it comes to matters like who reads what, outcomes arrived at freely are always better than outcomes achieved through compulsion.

(b) **Liberalism as Non-Compulsion** does not fall prey to Gibbard-style objections. It is easy to see that the principle on its own could never give rise to cycles of betterness. The set of all possible states-of-affairs is partitioned by those in which all self-regarding choices are free and those in which some self-regarding choice is unfree. And all Liberalism as Non-Compulsion says is that everything in the former set is better than everything in the latter set. On its own, this won't lead to a cycle.

14. However, conjoined with the **Weak Pareto Principle**, this new principle of liberalism does run into cycles, in precisely the same way as before.

(a) Let's return to Sen's example of Prude and Lewd. First, we'll distinguish between those possibilities in which all choices are free and those possibilities in which some choices are unfree. We'll subscript each of the previous states-of-affairs with an 'F' if they are states-of-affairs in which all choices are *free* and with a 'U' if it is a state-of-affairs in which people are compelled against their will to act in a certain way.

- N_F : Both Prude and Lewd freely choose to not read.
- N_U : Both Prude and Lewd are forced to not read.
- L_F : Lewd freely chooses to read and Prude freely chooses to not read.
- L_U : Lewd is forced to read, and Prude is forced to not read.
- P_F : Prude freely chooses to read and Lewd freely chooses to not read.
- P_U : Prude is forced to read and Lewd is forced to not read.
- B_F : Both Prude and Lewd freely choose to read.
- B_U : Both Prude and Lewd are forced to read.

(b) Now suppose that, while both Prude values freedom—so that, all else equal, he would rather have Lewd and himself choose freely than be compelled—he values it less than he does keeping the filth of *Fifty Shades* from spreading. And, while Lewd values freedom—so that, all else equal, he would rather have Prude and himself choose freely than be compelled—he values it less than Prude's disgust, and his delight, at the book's depravity.

So, if we ignore the subscripts, their preferences are the same as before, and otherwise, each of them prefers outcomes where choices are free.

	Prude	Lewd
1st	N_F	B_F
2nd	N_F	B_U
3rd	P_F	P_F
4th	P_U	P_U
5th	L_F	L_F
6th	L_U	L_U
7th	B_F	N_F
8th	B_U	N_U

(c) Now, by **Liberalism as Non-Compulsion**, the outcome where Lewd freely chooses to read and Prude freely refrains is better than the outcome where Prude is forced to read and Lewd to refrain.

$$L_F \succ_G P_U$$

(d) But both Prude and Lewd prefer P_U to L_F. So, by **Weak Pareto Principle**,

$$P_U \succ_G L_F$$

(e) And this contradicts **No Cycles**

$$L_F \succ_G P_U \succ_G L_F$$

15. In sum: Gibbard showed us that we ought to reformulate Sen's Minimal Liberalism. However, the best reformulation doesn't get us out of the conflict with the Weak Pareto Principle. So the conflict is genuine. If you are a liberal, you cannot be a Paretian. If you are a liberal, you should deny that people's preferences determine goodness in the way that Pareto imagined. If you are a Paretian, then you cannot be a liberal. If you are a Paretian, you should deny that it's always best for self-regarding decisions to be left to the individual.

(a) Of course, all of this assumes consequentialism. Another way to get out of the conflict is to accept the Pareto Principle but think that some actions are *wrong*, even if they would lead to *better* consequences. (This is the position of Nozick, for instance.)

10.2 *Égalité*

> *What is a social welfare function? Present the formal framework that Beckerman introduces for thinking about the goodness of states of affairs in terms of (the indifference curves of) social welfare functions and the Pareto Frontier. What does Harsanyi's theorem say? What is Egalitarianism, and what is the 'Leveling Down' objection to Egalitarianism? Explain how Beckerman's framework can be used to offer a response to the 'Leveling Down' objection to Egalitarianism.*

16. We saw in the previous section that we *cannot* incorporate a robust commitment to liberty into our social welfare functions without giving up the Pareto Principle. What about a commitment to *equality*?

10.2.1 Harsanyi's Theorem

17. A quick reminder:

(a) If we suppose that everyone's individual preference ordering $\succeq_i$ satisfies reflexivity, transitivity, and completeness, then we may represent these preference orderings with an *ordinal* utility function, U_i, and our social welfare function will be a function of these ordinal utilities,

$$\succeq_G = G(\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_i, \ldots \mathcal{U}_N)$$

(b) If we suppose further that everyone's preference ordering satisfies continuity, sweetening, better chances, and reduction of compound lotteries, then we may represent their preferences with a *cardinal* utility function, and our social welfare function will be a function of these *cardinal* utilities.

18. Harsanyi: what if, in addition to assuming that each *individual* has preferences representable with a cardinal utility function, we require that the *group* preference relation satisfy reflexivity, transitivity, completeness, continuity, sweetening, better chances, and reduction of compound lotteries?

    (a) Then, our group preference ordering, $\succeq_G$ would be representable as an expected utility-maximizing cardinal utility function, $\mathcal{U}_G$, which would just be a function, G, of the cardinal utility functions of the members of society.

    $$\mathcal{U}_G = G(\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_i, \ldots \mathcal{U}_N)$$

    (b) What if we *further* require our social welfare function to satisfy the following principle:

    Strong Pareto
    For any two states S and T:
    (1) If everyone is indifferent between S and T, then S is equally as good as T
    (2) If someone strictly prefers S to T and no one strictly prefers T to S, then S is better than T.

    i. (2) is just the regular Pareto Principle. So **Strong Pareto** entails the regular Pareto Principle.
    ii. The extra strength of the strong principle comes from (1).

    (c) Harsanyi shows that, if the individual's preferences are representable as cardinal utility functions, the social preferences determined by a social welfare function are representable as expected utility-maximizing cardinal utility functions, and the social preferences satisfy the Strong Pareto Principle, then the group utility function $\mathcal{U}_G$ is just a *weighted sum* of the utilities of all of the individuals in society.

    Harsanyi's Theorem
    If our social welfare function G satisfies the **Strong Pareto** principle, each individual's $\succeq_i$ is representable as a cardinal utility function, and the group preference determined by G, $\succeq_G$, is representable as an expected utility-maximizing cardinal utility function $\mathcal{U}_G$, then there are some $\alpha_1, \alpha_2, \ldots, \alpha_N$ such that, for each state S,

    $$\mathcal{U}_G(S) = \alpha_1\mathcal{U}_1(S) + \alpha_2\mathcal{U}_2(S) + \cdots + \alpha_i\mathcal{U}_i(S) + \ldots \alpha_N\mathcal{U}_N(S)$$

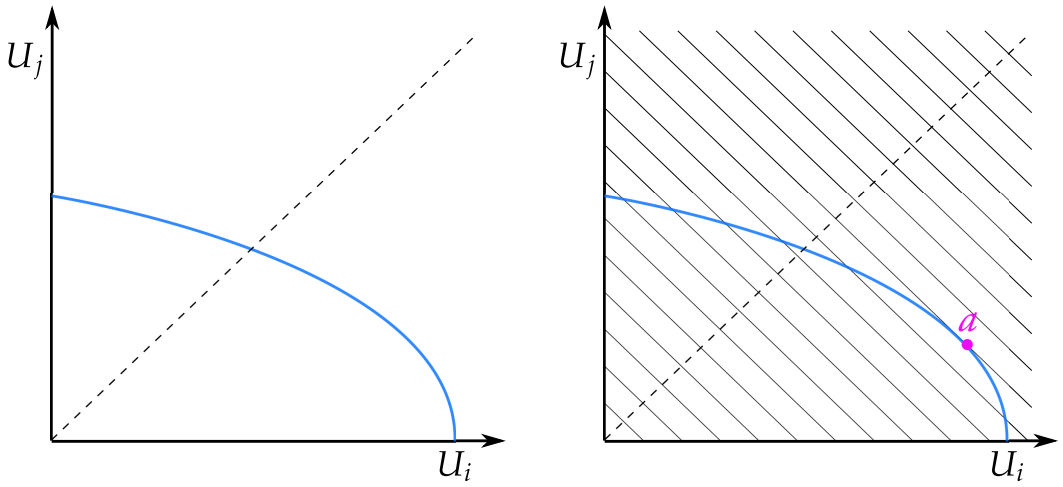
    (d) If we additionally constrain every individual's utility function to be bounded between 0 and 1—by adopting the zero-one rule—then we can say something stronger: namely, that the group utility function is just the sum of the individual's utility functions.

    Harsanyi's Theorem (v2)
    If our social welfare function G satisfies the **Strong Pareto** principle, each individual's $\succeq_i$ is representable as a cardinal utility function whose minimum value is 0 and whose maximum value is 1, the group preference determined by G, $\succeq_G$, is representable as an expected utility-maximizing cardinal utility function $\mathcal{U}_G$, then, for each state S,

    $$\mathcal{U}_G(S) = \mathcal{U}_1^{z-o}(S) + \mathcal{U}_2^{z-o}(S) + \cdots + \mathcal{U}_i^{z-o}(S) + \ldots \mathcal{U}_N^{z-o}(S)$$

19. We can get a nice visual understanding of what's going on here by looking at the indifference curves of this linear social welfare function.

    (a) Assume that we have a common scale on which to make interpersonal comparisons of utility (the zero-one rule will do, for instance).



(a) (b)

Figure 10.1: Figure 10.1a shows the Pareto frontier. The dotted 45° line is the line of equal utility for Ike and Janet. Figure 11.2 adds the indifference curves of a linear social welfare function. With this social welfare function and this Pareto frontier, point a is the optimal state-of-affairs.

(b) There will be various utility packages which are attainable given the current state of the economy; these attainable utility packages may be thought of on analogy with the various consumption bundles allowable given budget constraints.

i. For simplicity, consider just two people, Ike and Janet, and let Ike's utility function lie along the x-axis; and Janet's utility function lie along the y-axis, as in Figure 10.1a.

ii. Then, the state of the economy will determine a set of possible allocations of utility to Ike and Janet.

iii. The border of this set will be the *Pareto frontier*, which contains all of those Pareto optimal allocations of utility to Ike and Janet.

(c) We then may draw the *indifference curves* of the *social welfare function*, G. Supposing that we have a linear social welfare function of the kind from Harsayni's theorem, we will get the indifference curves shown in figure 11.2.

(d) The point on the Pareto Frontier which lies tangent to these indifference curves will then be the point at which group welfare is maximized. Assuming welfarism, this will be the optimal allocation. Assuming consequentialism, this is the allocation we should adopt.

10.2.2 Prioritarian Social Welfare Functions

20. Harsanyi's linear social welfare function is sensitive only to the total amount of utility, it is indifferent to questions of distribution. It regards Ike having a utility of 1 and Janet having a utility of zero as being just as good as Ike and Janet both having a utility of 0.5.

(a) Thus, it can deem optimal distributions of utility which are quite far from the line of equal distribution (see figure 11.2).

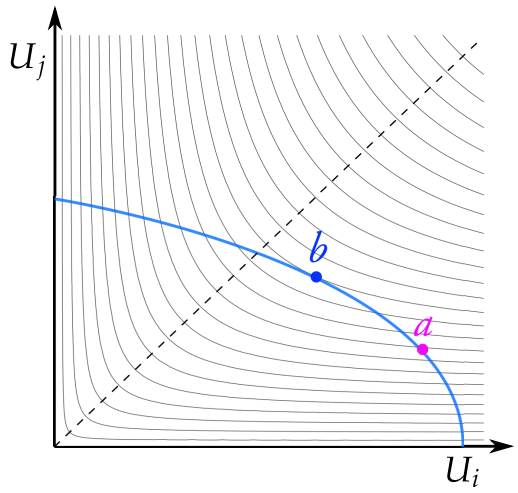


Figure 10.2: The Pareto frontier, along with the indifference curves of a prioritarian welfare function. With this social welfare function and Pareto frontier, point b is the optimal state-of-affairs.

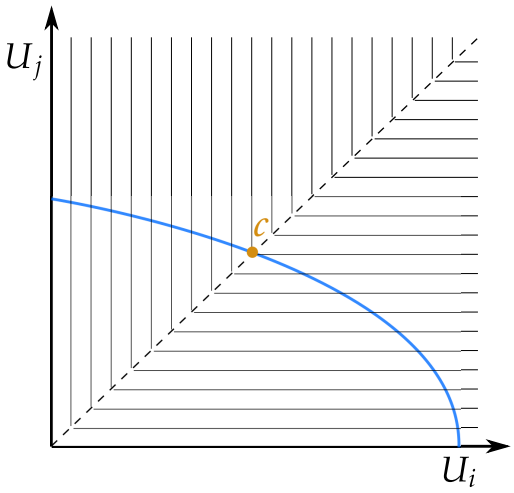


Figure 10.3: The Pareto Frontier, along with the indifference curves of a Rawlsian, or maximin, social welfare function. With this social welfare function and Pareto frontier, point c is the optimal state-of-affairs.

(b) Recall Hausman's objection to preferentism: this will tell us that we ought to cater to those with expensive tastes..

21. If we object to these kinds of ethical consequences—and, in particular, if we think that inequality makes a state of affairs worse—then we will want to adopt a different social welfare function.

    (a) One option would be to say that the *weight* attached to an individual's welfare is greater the worse off they are, and lesser the better off they are.

    (b) Suppose that, in the case of Ike and Janet, we decide that, for every state-of-affairs S,

    $$\mathcal{U}_G(S) = \left(\frac{\mathcal{U}_j(S)}{\mathcal{U}_i(S)+\mathcal{U}_j(S)}\right)\cdot\mathcal{U}_i(S) \;+\; \left(\frac{\mathcal{U}_i(S)}{\mathcal{U}_i(S)+\mathcal{U}_j(S)}\right)\cdot\mathcal{U}_j(S)$$

    So that, as Ike becomes worse off, compared to Janet, his utility becomes *more* important; and, as Ike becomes better off, compared to Janet, his utility becomes *less* important.

    (c) This is a kind of *prioritarian* social welfare function—so named because it gives *priority* to the least well off. It produces the indifference curves shown in figure 10.2.

    (d) Bear in mind: Harsanyi's theorem tells us that $\mathcal{U}_G$ will not be an expected utility-maximizing cardinal utility function.

22. Notice that the prioritarian social welfare function ends up favoring more equal distributions of utility; though it doesn't achieve this by valuing equality *per se*. Rather, it just values the utility of the *worst off* more, and this ends up having the consequence that more equal distributions of utility are better.

23. One particularly radical kind of prioritarian social welfare function is the one which says that it is *only* the utility level of the least well-off person which matters.

(a) John Rawls advocated a social welfare function roughly like this. We can call it a 'Rawlsian', or the 'maximin' social welfare function ('maximin' because it attempts to *maximize* the *minimum* utility). It says that the group utility is just equal to the least utility of any member of society. That is, for any state of affairs S,

$$\mathcal{U}_G(S) = \min\{\mathcal{U}_1(S), \mathcal{U}_2(S), \ldots \mathcal{U}_i(S), \ldots, \mathcal{U}_N(S)\}$$

(b) If we confine our attention to just Ike and Janet, then $\mathcal{U}_G(S) = \min\{\mathcal{U}_i(S), \mathcal{U}_j(S)\}$. The indifference curves of this social welfare function are shown in figure 10.3.

(c) In the static case, the Rawlsian social welfare function will always require perfect equality. However, in the dynamic case, it may allow inequality, so long as this inequality produces greater utility for the least-well-off later on.

10.2.3 The 'Leveling Down' Objection

24. If we value equality any more than the Rawlsian social welfare function does, then we will end up violating the Pareto Principle.

    (a) Consider the *very* egalitarian social welfare function in figure 10.4.

    (b) That social welfare function says that e is better than d.

    (c) However, the Pareto Principle tells us that d is better than e—for Ike prefers d to e, and Janet is indifferent between d and e.

25. This conflict between the *very* egalitarian social welfare function and the Pareto Principle is one way of understanding the 'leveling down' objection to egalitarianism.

    (a) If we were at d, then it would be better to 'level down' to e—to make Ike worse off without making Janet any better off.

    (b) Notice, though, that if this is the 'leveling down' objection to egalitarianism, then it isn't a very good objection. The egalitarian can just endorse one of the social welfare functions shown in figures 10.2 or 10.3, and they will value equality without violating the Pareto Principle, and without thinking that leveling down is ever an improvement.

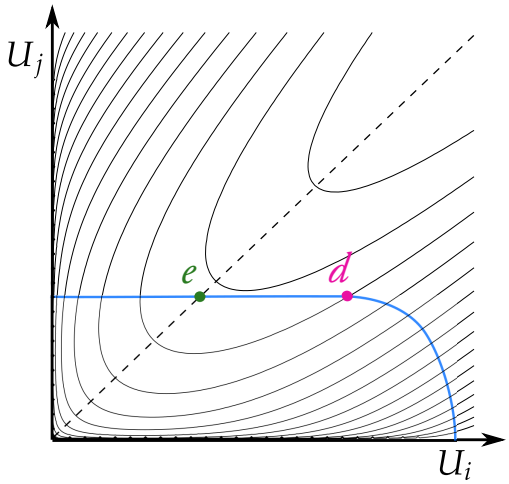


Figure 10.4: The Pareto Frontier, along with the indifference curves of a *very* egalitarian welfare function. This egalitarian social welfare function violates the Pareto Principle. Ike prefers d to e, and Janet is indifferent between d and e, yet this social welfare function tells us that e is better than d.

Chapter 11

Valuing Lives

11.1 Valuing Life

1. Many of society's choices seem to require us to evaluate the goodness/badness of outcomes involving *loss* of life or various *risks* to life. (Alternatively, if we are not consequentialists: the rightness/wrongness of the *actions* which bring about the loss of life or the risk of loss of life.)

   (a) When it comes to health care, there are only so many resources available. Giving a kidney to one means depriving another of that kidney. What should hospitals do?

   (b) Projects like SpaceX involve substantial risk to life for some, with potential benefits for others down the line. What kinds of risks should people be subjected to, and for what ends?

   (c) Restricting carbon emissions in developing countries leads to less economic growth, but also improved life and health outcomes for all down the line. How should these considerations be balanced?

2. Several ethical theories we have already studied provide at least partial answers to these questions.

   (a) Utilitarianism says that we should look at the total pleasure which would (likely) be brought about by each course of action, and use it to decide.

      i. For the utilitarian, life itself has no intrinsic value; it is only instrumentally valuable insofar as living allows people to experience pleasure.

         A. The kidney should be given to the person who would enjoy it most, for the longest period of time.

         B. SpaceX should submit people to risks iff the future pleasure which will (likely) result from those risks outweighs the loss of pleasure (in expectation) to those who may die.

      ii. Recall that the disregard Utilitarianism pays human life was one of our primary criticisms of that view. Utilitarianism says that it is permissible to kill a person so as to harvest their organs, so long as those who receive the organs are happier for longer than the killed person would have been.

   (b) Kant's moral theory says that it is impermissible to use others as mere means to your own ends. You may enjoin others to help in your projects only if they have freely and informedly

consented to their involvement. Your decisions should follow from a maxim (or rule) which you could consistently will to be universally followed.

i. If a policy for distributing kidneys is one you can consistently will to be universally followed, then it is a permissible policy for distributing kidneys.

A. Note that, if the utilitarian's policy is consistently universalizable in both will and conception, then the Kantian can accept that it is permissible.

ii. If those aboard the SpaceX rocket have freely consented to the risks involved, then Kant's moral theory says that those risks are permissible.

iii. Since those harmed by carbon emissions have not freely consented to those harms, any policy which jeopardizes their well-being for the good of others is not permissible. It is using those harmed as mere means, rather than as ends in-and-of-themselves.

(c) Libertarianism claims that individuals have a claim right, against others, for them to not deprive you of life without your consent.

i. So long as nobody has a claim right *for* a kidney, any policy of distribution is just (including, perhaps, a utilitarian one).

ii. Individuals have the power right to waive their claim right to not have others expose them to undue risk of harm, or death, by freely and informedly consenting to the risk. So the risks involved in SpaceX are permissible.

iii. What does libertarianism say about carbon emissions and climate change? One possible answer: Nozick says that property may be acquired only if, in acquiring it, you do not violate the rights of others. If anybody in Bangladesh has a right to the use of their land, then acquiring goods through emitting gross amounts of carbon involves rights violations. [But perhaps there is room to quibble here.]

3. Note that many of these approaches—Kantianism and Libertarianism, in particular—contend that the way to make decisions involving the potential loss of life is *not* to find some way of valuing that life, and then calculating whether the loss of life is worth the benefits.

4. Beckerman calls these views 'heroic', and argues against them as follows (this is my best reconstruction of the argument):

P1. We cannot justify spending all of GDP in order to guarantee that people only die of old age.

P2. If human life is sacred and inviolable, then we would have to spend all of GDP in order to guarantee that people only die of old age.

C1. Human life is not sacred and inviolable. [from P1 and P2]

C2. We must use valuations of human life in order to decide where to allocate scarce resources. [from C1]

(a) It is unclear what Beckerman means by 'sacred and inviolable'. If he means 'humans have a right to life', then P2 is clearly false, as the examples of Kantianism and Libertarianism makes clear. Just because humans may not be killed or exposed to great risk of death (without consent) for some greater good, this does not mean that we have an obligation to devote as many resources as possible towards keeping them alive.

(b) Even if P2 is granted, C2 does not follow from C1. Even granting that we must make decisions about how to allocate scarce resources like kidneys, unless we assume a particular kind of consequentialism, we needn't think that the only way of doing this is by calculating the value of the human life.

i. For instance, we could use considerations of *fairness* to decide who should be given access to scarce kidneys, without ever having to say how valuable an individual life is.

ii. Additionally, even if we are consequentialists, we needn't think that the goodness of a state of affairs is determined by calculating the value of the human lives which exist in those states of affairs.

5. Beckerman's argument against alternative approaches is unconvincing. However, if we accept a general consequentialist position like his, we will need some way of valuing (risks to) human lives.

11.1.1 Kaldor Hicks and Valuing Life

6. Recall the Kaldor Hicks approach to evaluating policy proposals:

(a) Some policies will have *winners* and *losers*. If the winners *could* compensate the losers in such a way that the policy would constitute a Pareto improvement, then Kaldor Hicks says that implementing the policy is an improvement on the current state-of-affairs.

▷ Importantly, Kaldor Hicks does *not* require that the winners *actually do* compensate the losers.

(b) Recall: these evaluations are not anti-symmetric. Kaldor Hicks can say that S is an improvement on T, and also that T is an improvement on S.

7. Beckerman suggests that the Kaldor Hicks approach will be of no help when determining how to evaluate policies which will result in some dying.

(a) For illustration: suppose that there is a construction project which will bring safe drinking water to many who lack it, but will result in a few dying during construction.

(b) With this project, there will be winners (those with clean and safe drinking water), but they will not be able to compensate the losers, since the losers are no longer alive.

(c) Even in principle, there's no amount of money the losers would be willing to accept in exchange for their loss of life.

(d) So, Beckerman concludes: Kaldor Hicks will not help us evaluate project like these.

11.1.2 The 'Net Output' Method of Valuing Life

8. Beckerman next considers the 'net output' method of valuing a life.

9. On this approach, we ask how much value the human life has for the rest of society.

(a) That is: the dollar value of a life (during some time period) is given by the total amount that person contributes to GDP during that time period.

(b) And: the dollar *dis*value of a loss of life is given by the difference in GDP between the contingency in which that person lives and the GDP in which that person dies.

10. This approach has some morally monstrous consequences.

    (a) According to it, the lives of those who contribute more to society are worth more.

    (b) Worse: those who live on government assistance will contribute *negative* dollars. This means that, according to the net output method, it would be an improvement if they died.

11.1.3 The 'Willingness to Pay' Method of Valuing Life

11. The 'Willingness to Pay' approach evaluates risks to human life by asking how much people would require to assume those risks (or, how much they would be willing to pay to alleviate those risks).

12. For example, suppose that there is a construction project which will bring safe drinking water to many who lack it. For each of the 1,000 people involved in the project, there is a 0.1% chance they will die during construction.

    (a) Thus: we expect 1 person to die, though we're not sure how it is.

    (b) We provide people with surveys, and ask them how much money they would require in order to assume a 0.1% chance of dying. Suppose they say that they would ask for $100.

    (c) Then, we calculate the cost of the construction to be $\$100 \times 1{,}000 = \$100{,}000$. If the benefits are more valuable than the costs, then we should do the project.

13. John Broome criticizes the WTP approach because it has the implausible consequence that our knowledge of who will die makes a difference with respect to whether the project is worth pursuing.

    (a) For illustration: suppose we discover *who exactly* will die in the project. At that point, the risk this person faces changes from 0.1% to close to 100%. They will now demand more money for their involvement with the project. Suppose they demand $10,000,000. Then, according to WTP, the cost of the project has risen from $100,000 to $10,000,000.

    (b) But, Broome contends, whether we know who it is that will die should not make a difference with respect to the question of how valuable their lives are. WTP says that this knowledge *does* make a difference. So, WTP gives bad advice about how to measure the value of human life.

11.2 The Non-Identity Problem

14. Consider Parfit's story of *The 14-Year-Old Girl*:

> This girl chooses to have a child. Because she is so young, she gives her child a bad start in life. Though this will have bad effects throughout the child's life, his life will, predictably, be worth living. If she had waited for several years, she would have had a different child, to whom she would have given a better start in life.

(a) A natural reaction to the case is that she has made the wrong decision—not only for *herself*, but also for *her child*. It is natural to wish to tell the 14-year-old girl, that choosing to have a child at such a young age is harming *her child*. Had she waited, she would have given her child a better life.

(b) However, Parfit thinks that this is simply not true. Rather, she would have had a *different* child, had she waited. He relies upon the following assumption:

The Time-Dependence Claim If any particular person had not been conceived when they were in fact conceived, then they never would have existed.

i. Suppose, for instance, that, in order for *you* to exist, somebody with your genetic code must exist. Then, if you hadn't been conceived when you in fact were, somebody with a *different* genetic code would have existed, and that person wouldn't have been *you*.

(c) Yet Parfit thinks that it was *worse* for the 14-year-old girl to have the child.

15. Another case with a similar structure: we must choose whether to conserve or deplete the world's resources. If we choose depletion, then the well-being of the people who live over the next 200 years will be slightly higher, but the well-being of people forever after will be much lower (though their lives will still be worth living).

(a) A natural objection to depletion is that it is worse *for future generations*.

(b) However, again, Parfit thinks that this is simply not true. If we choose depletion, a great many things about the history of humanity will be different. These changes will make a large difference with respect to which people end up living. In fact, if we choose depletion, then nobody who will be around in 200 years time *would* have been around, had we chosen conservation.

(c) We could make the following argument for the conclusion that depletion is *not* worse than depletion:

P1. If S is worse than T, then S is worse than T *for somebody*.

P2. There is nobody for whom depletion is worse than conservation.

C1. So: depletion is not worse than conservation.

16. The conclusion of this argument is absurd. Since P2 is true, Parfit rejects P1. He concludes that things can be worse overall, without them being worse *for anybody*.

17. This raises a question: how do we compare states-of-affairs when there are different people around in those states-of-affairs?

11.3 The Repugnant Conclusion

18. A preview: Parfit is going to think about how we should compare the goodness of states where different people exist. He's going to make some apparently mundane claims, which will lead him to a *repugnant conclusion*.

19. Consider the state, A, shown in figure 11.1a. There, the width of the rectangle represents the number of people who are alive, and the height of the rectangle represents their level of well-being. In A, there are a few people living very good lives.

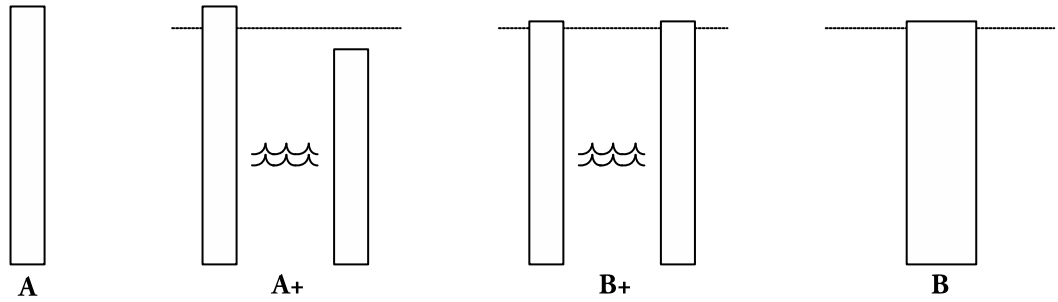


Figure 11.1: The width of the rectangles represents the number of people. The height represents the level of well-being of those people. The dotted line in figures 11.1b, 11.1c, and 11.1d indicates the *average* level of well-being of people in the state $A+$.

20. Consider now the state, $A+$, shown in figure 11.1b. This state is just like A, except that we have added to it some additional people, separated from those in A by an ocean. These new people have lives which are worth living, though they are not *quite* as well off as the original people from A.

    ▷ Parfit's first claim: $A+$ is not worse than A. Getting from A to $A+$ involves the 'mere addition' of some people whose lives are worth living. This cannot make things worse. There is no injustice in $A+$, since these additional people are separated from the original people from A.

21. Consider now the state, $B+$, shown in figure 11.1c. This state is just like $A+$, except that the well-being of the original people from A has been slightly lowered, and the well-being of the new people on the right has been slightly raised. The *average* level of well-being has, however, been raised.

    ▷ Parfit's second claim: $B+$ is better than $A+$. It is better if there is less inequality in the world; and, moreover, since the *average* level of welfare has gone up, so too has the *total* level of welfare. So even those who care not at all for equality and simply want to maximize total welfare will think that $B+$ is better than $A+$

    Finally, consider the state B, shown in figure 11.1d. Here, we have merely brought together the people from the island and the original people from A.

    ▷ Parfit's third claim: B is just as good as $B+$. We've not changed anybody's level of wellbeing. We've changed only their spatial location. Moreover, there is no injustice, since everybody has the same level of wellbeing.

22. Putting together Parfit's claims thus far, we can argue that B is better than A.

    P1. $A+$ is not worse than A

    P2. $B+$ is better than $A+$

    P3. B is as good as $B+$

    C. B is better than A.

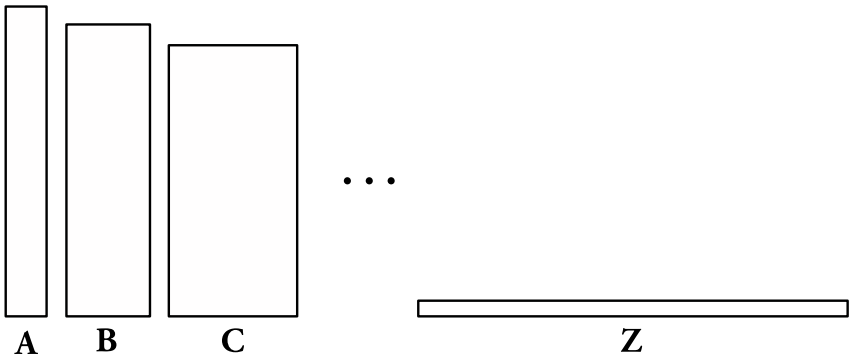


Figure 11.2

(a) This is the conclusion for which Parfit argues—but we could aim for a slightly weaker conclusion: B is not worse than A. This will also lead us to a troubling conclusion.

23. What we've seen is that, so long as people have lives worth living, we can *decrease* the average level of wellbeing by simply *increasing* the number of people. But by repeating such modifications over and over again (see figure 11.2), we can reach a state in which there are *tons and tons* of people, all living lives *barely* worth living (lives of 'potatoes and muzak', as Parfit puts it).

(a) Using the same reasoning as before, we can argue for each of the premises below:

B is better than A	B is not worse than A
C is better than B	C is not worse than B
$\vdots$	$\vdots$
Y is better than Z	Y is not worse than Z
Z is better than A	Z is not worse than A

24. But these conclusions, Parfit submits, are *repugnant*. Surely it is better for fewer people to exist with *very* high levels of wellbeing than for *tons and tons* of people to exist and live lives barely worth living.

25. So, how, in general, should we value states of affairs involving different people? Let's consider some proposals and think about what they say about Parfit's *repugnant conclusion*.

26. One proposal:

The Total Principle The goodness of a state of affairs is given by summing up the *total amount* of welfare of all of the individuals existing in that state of affairs.

(a) If the total principle is correct, then our reasoning above was flawless. B *really is* better than A, C *really is* better than B, and so on.

(b) We are driven straight to the repugnant conclusion. Z is better than A.

27. Perhaps, then we shouldn't look at the *total* amount of welfare. Perhaps, instead, we should look at the *average* level of welfare.

    The Average Principle The goodness of a state of affairs is given by the *average* level of well-being of all of the individuals existing in that state of affairs.

    (a) The average principle allows us to escape the repugnant conclusion. How does it do this? By denying the first premise: it denies that $A+$ is not worse than A.

    (b) According to the average principle, merely adding additional people with lives well worth living can make matters *worse*.

28. The average principle escapes the repugnant conclusion, but it faces many other objections.

    (a) If your level of wellbeing is slightly less than average, then things would have been better had you not been born.

        ▷ If we additionally assume consequentialism: it would be right to painlessly kill off the least happy.

    (b) Suppose you are deciding whether to have a child, and you know what it's level of wellbeing would be—call it 'w'. According to the average principle, whether having the child would be good or not can depend upon the welfare of people who lived in the distant past or far away.

        ▷ For suppose that there were many people in the past/far away with welfare much higher than w. Then, having a child could make things worse.

        ▷ On the other hand, if were were many people in the past/far away with welfare much *less* than w, then having a child could make things better.

    (c) The average principle also entails what we can call *the sadistic conclusion*. Compare the following two states-of-affairs:

        $A+$ As in figure 11.1b, on a distant island, a large number of people living very good lives (but lives *slightly* less worth living than those from A) come into existence.

        $A-$ Everybody from A exists, but, on a distant island, a *single* person comes into existence, and is tortured for their entire life.

        So long as the suffering of the one tortured person brings down the *average* welfare less than the many people with lives worth living, the average principle says that $A-$ is *better* than $A+$.