Core Prelims lectures on Moral Philosophy (via Mill's *Utilitarianism*)

Michaelmas Term 2013

Dr Hilary Greaves

Lecture 1: Theories of well-being I: Hedonism

0. Admin
    a. These are the core lectures for the Prelims/Mods option in 'Moral Philosophy'. Edward Harcourt will offer a second set of core lectures for the same option next term. Each is intended to be a 'complete' introduction to the syllabus material, but they are not duplicates (you can choose to go to either or both sets of lectures).
    b. The Faculty reading list for this paper is available on Weblearn (philosophy website -> undergraduate -> reading lists -> Mods and prelims reading lists -> Moral philosophy reading list).
    c. References to Mill's Utilitarianism are given in the form 'U [Chapter].[paragraph]': e.g. 'U 2.10' refers to the 10th paragraph of chapter 2.
1. Some remarks on the context, and lecture outline
    a. The basic questions of (normative) ethics are "What should I do?" and "How should I live?"
    b. Utilitarianism is an attempt to answer these questions.
        i. Its answer places great emphasis on the notions of happiness/pleasure, and unhappiness/pain.
        ii. Example: I ought to give the best set of lectures on Mill that I'm able to prepare in the time available.
            1. But why ought I to do this? A shallow (?) answer might just say: As students admitted to (and paying fees to!) Oxford University, you have a *right* to an excellent education, and as a lecturer here I have a *duty* to provide you with that.
            2. A utilitarian would not be satisfied with this sort of answer. She might be thinking instead: It's good if you lot understand Mill and utilitarian theory as well as possible. It's good if the lectures help you to develop your critical skills, by providing a good example. And it's good if you enjoy the lectures, so I shouldn't make them too boring.
            3. For the first of those two 'good things' – understanding utilitarianism and developing critical skills – we might go on to ask, in turn, why *those* things are good.
                a. It's good if you understand utilitarianism so that in your private lives, and in your future careers, you're more likely to act in ways that make other people happy. Also, properly understanding such a beautiful theory is itself an exquisite intellectual pleasure for you, even aside from its consequences on your action.

b. It's good to develop your critical thinking skills, because that makes you more effective in whatever you try to do – and, hopefully, most things you try to do will be things that make yourselves and others happier...
    i. Exception: If you were a budding master criminal, then it *wouldn't* be good to hone your critical thinking skills!
c. Utilitarians are often driven partly by the thought that one can always ask *why* something is good, unless the thing in question is happiness/pleasure/enjoyment.
    i. If it isn't itself happiness, the thing is (according to utilitarians!) good at most in an *instrumental* sense – it has (at most) *instrumental value.* That is, its goodness consists in its tendency to bring about some further thing that's good, and Thing 1 is good *only insofar as* Thing 2 is good (and only insofar as Thing 1 really does tend to bring about Thing 2).
    ii. Utilitarians think that
        1. all such 'chains of instrumental goods' end eventually at instances of happiness/pleasure/enjoyment;
        2. there's no answer to be given to "why is happiness good" other than "it just is". Happiness has *final value* - so this is where the chains of 'explaining why something is good' stop.
4. How these lines of thoughts relate to considerations of rights/duties is an interesting question – of which more later!

iii. The idea that what one ought to do is intimately related to considerations of the "pleasure and pain" or "happiness and unhappiness" that would result from one's action is not a new one. Indeed, every remotely sane theory of ethics, throughout history, has counted this *at least one part of* the story. What's distinctive about its answer is that it reduces questions of what to do/how to live *entirely* to empirical questions of what will maximise happiness/pleasure, and minimise unhappiness/pain.
iv. This is both the feature that draws utilitarianism's supporters to the theory, and the feature that objectors find to be (on reflection) totally unacceptable.
    1. Your task: Figure out where you stand on this!

c. We will study the theory of utilitarianism primarily via the work of John Stuart Mill.
    i. Mill was a 19th century philosopher, political theorist, economist and civil servant. (1806-1873)
    ii. His father was James Mill, a close friend and associate of the famous utilitarian Jeremy Bentham.

iii. James Mill brought his son up with the deliberate aim of creating a "genius intellect" that would carry on Bentham's work of developing and advocating for utilitarianism.
iv. This involved an extremely rigorous programme of home schooling, including beginning studies in Greek at the age of three. By the age of fourteen, Mill was attending university courses in chemistry, zoology, logic and higher mathematics, in France, and associating with many of his father's political allies.
v. But it seems to have worked!
    1. Mill started publishing writings on ethical theory and practice at the age of fourteen.
    2. *Utilitarianism* was first published in 1861 (when Mill was 55).
    3. Mill's writing on utilitarianism is prompted by criticisms that Bentham's version of utilitarianism had received.
vi. Utilitarianism is intended for a *general audience*. (It was originally published in serial form, in *Fraser* magazine, and only afterwards reprinted in book form.) This is worth bearing in mind when you read it. It is not, for example, anything like as clear as one would like, on several points of detail that are crucial to the serious student of utilitarianism; and, as we'll see, some of Mill's moves are made for rhetorical effect, rather than as presentations of a carefully considered case. It is rather focussed on defending the general idea and emphasising the big picture.
    1. Scholars continue to this day to debate how Mill's *Utilitarianism* ought to be interpreted. (In your tutorials, you might look into some of these debates.)
    2. For our purposes, Mill's vagueness on these points is actually rather helpful: we will use his book as a jumping-off point to investigate the alternatives, and think about what the most plausible version of utilitarianism is.

d. Outline of these lectures
    i. Week 1: Theories of well-being I: Hedonism
    ii. Week 2: Theories of well-beingII: Desire-satisfaction and objective-list theories
    iii. Week 3: Theories of the good (utilitarianism and alternatives); Criteria of right action I: Act-consequentialism
    iv. Week 4: Criteria of right action II: Rule-consequentialism, Mill's account, non-consequentialist theories
    v. Week 5: Meta-ethics I: Mill on moral motivation and moral epistemology
    vi. Week 6: Meta-ethics II: Meta-ethical realism and alternatives
    vii. Week 7: Mill on justice and rights
    viii. Week 8: Two-level consequentialism and global consequentialism

2. 'Theories of well-being': Candidates answers to the question 'What makes a person's life go well *for her*?'
    a. In particular: Is it really only *happiness* that (ultimately) matters, for that purpose?

i. A positive answer to this question is the first thing that distinguishes utilitarians from many of their opponents.

b. Clarification of the question
   i. There is *one* clear sense in which a person's life uncontroversially '*goes well*' if that person does great things for others: e.g. discovers a cure for cancer, brings universal primary education to the populace of an impoverished country, abolishes slavery, etc. That sense is: this person's life *improves the overall state of the world*, so it is *a good thing* that this person lived.
   ii. But such a life might not be good *for the person living it*. It might just be good *for other people*. (Suppose, e.g., that the person *never found out* that he had discovered the cure for cancer, that he did not himself either enjoy or see any value in his research, that his work was not appreciated in his own lifetime – indeed, suppose that by some terrible misunderstanding his work led to him being widely ridiculed and even punished during his lifetime. In that case, it would at least be *highly controversial* to claim that his life was good *for him*.)
   iii. We need to have an account of what it is for someone's life to be good *for them* before we are in a position to consider what it is for someone's life to go well overall (since we don't know how good it is, e.g., that a cure for cancer is discovered, or that a beautiful artwork is produced, unless we know to what extent this improves the lives of the people who benefit from it – for instance, how much *health* and *length of life* contribute to well-being.)

c. Hedonism: What's good for a person is for her life to contain as much happiness and as little unhappiness as possible.
   i. Motivations
      1. Similar to the above thought-experiment: Choose any action you find yourself doing *for non-moral reasons* – reasons that we might call 'selfish', in a non-derogatory sense – and that, on reflection, you still think is a good idea. Ask yourself: why are you doing it? In most (all?) such cases, you can find a plausible line of 'why'-answers that ends with: "because I want to be happy/I don't want to be unhappy". And that appears (?) to be where the line of explanation stops. (Try it!)
      2. "Pleasure, and freedom from pain, are the only things desirable as ends; and... all desirable things... are desirable either for the pleasure inherent in themselves, or as means to the promotion of pleasure and the prevention of pain." (U 2.2)
   ii. Versions of hedonism
      1. 'Caricature' hedonism ('sensualism'): Only bodily pleasures and pains count
         a. The word 'hedonism' in popular culture often connotes this version of hedonism: it's all about food, drink, sex, massages from gorgeous and scantily clad guys/girls...

    b. Not a straw man! “Fill your belly. Day and night make merry. Let days be full of joy. Dance and make music day and night [...] These things alone are the concern of men” (from the *Epic of Gilgamesh*, c.1800 BCE))
2. Making hedonism more plausible: Include intellectual and other mental pleasures. Examples:
    a. Reading poetry
    b. Social pleasures: the pleasures of good conversation, and the feelings of true friendship/romantic love
    c. Creative pleasures: composing or playing music, programming computers, making handicrafts
    d. A sense of satisfaction: e.g. at writing a book or climbing a mountain
3. Few people would really be satisfied – “happy” – with the life recommended by crude hedonism.
    a. Hedonism (and hence utilitarianism) has sometimes been criticises as “a doctrine worthy only of swine”: one that ignores the fact that the ways in which a human life can be good far outstrip the ‘mere pleasures’ to which other animals are susceptible.
    b. But this criticism applies only to an overly crude, sensualist version of hedonism.
4. Bentham’s “hedonic calculus”
    a. For every pleasure and every pain, note its *duration*, and its *intensity* (how pleasurable or how painful it is). Work out the value of a pleasure by multiplying its duration and its intensity; append a minus sign when dealing with pains. Then add up all the pleasures and pains that a given action will cause, to work out the overall value of the action.
    b. NB This includes *all* pleasures and pains, not only the ‘sensual’ ones
    c. Example: Climbing a mountain
        i. Suppose that you don’t actually enjoy the *process* of climbing the mountain, *at the time* – you’re only in it for the sense of achievement. To work out whether or not it is a good idea for you to climb a given mountain, work out how long it would take you to climb it, and what the intensity of the displeasure (‘pain’) that you would experience during the climb. (If the level of displeasure varies during the climb, split the journey into several bits, account for each separately and then add the results.) This is the total cost of climbing the mountain. Then estimate how intense the pleasure is that you would get from a sense of satisfaction at

having climbed it. Crucially, remember to take account of how long that lasts, too. If there are other reasons for the climb, e.g. you would enjoy chatting in the pub with your mates more if you had some mountain-climbing stories to boast about, then count the duration and intensity of those, too; if there are further knock-on effects due to your friends respecting you more, count them too. You should climb the mountain only if the total amount by which it increases your pleasure level is greater than the total amount by which it increases your pain or 'displeasure' level.

d. Note that the recommendation needn't be that people actually carry out this calculation, for every decision they take.
    i. A good thing, since such calculations would be horrifically complex.
    ii. Rather, it provides something to *aim* at, and something that one's decision-making reasoning might hope to *approximate*.
    iii. An exercise: Think through what Bentham's "hedonic calculus" would recommend in a particular example of an action you have chosen to do, or not to do. On reflection, do you think the recommendations of the hedonic calculus are reasonable?
e. This calculus *can* explain why intellectual pleasures are more valuable than more mundane pleasures – when they are!
    i. To some people, at any rate, they simply *provide more pleasure* – because the pleasure is more intense, because it lasts longer, or both.
    ii. They're relatively cheap: it doesn't require much outlay of pain to deliver them to a large number of people (in contrast to, e.g., fine dinners).
    iii. Relatedly: It's harder for unfortunate external circumstances to take them away.
        1. One might fall on hard times and not be able to afford material luxuries. But no ordinary misfortune (but cf. dementia, etc.) can rob one of one's sense of satisfaction, or enjoyment of contemplation.
f. But also, Bentham insists that intellectual pleasures aren't always more valuable, and that it's mere snobbery to insist otherwise. The ultimate tribunal is which activity provides the more pleasure.

i. "The utility of all these arts and sciences,-I speak both of those of amusement and curiosity,-the value which they possess, is exactly in proportion to the pleasure they yield. ... Prejudice apart, the game of push-pin is of equal value with the arts and sciences of music and poetry. If the game of push-pin furnish more pleasure, it is more valuable than either. Everybody can play at push-pin: poetry and music are relished only by a few. ... If poetry and music deserve to be preferred before a game of push-pin, it must be because they are calculated to gratify those individuals who are most difficult to be pleased." (Bentham, *The rationale of reward*, 1825)

d. Mill's version of hedonism: the distinction between 'higher' and 'lower' pleasures
    i. Mill notes the possibility, and the sufficiency, of a Benthamite defence of intellectual pleasures, but proposes instead to "take the higher ground" – to offer a defence of the superiority of higher pleasures according to which that superiority is more of a fundamental affair, not simply a result of contingent facts about which pleasures happen to be cheaper and longer-lasting.
    ii. And there's independent reason to expect this "higher-ground" defence to be correct: "It would be absurd to suggest that while, in estimating all other things, quality is considered as well as quantity, the estimation of pleasures should be supposed to depend on quantity alone." (U 2.4)
    iii. Some pleasures are *just more valuable* ("higher") than others.
        1. So, according to Mill, poetry is more valuable than pushpin, *even if* the quantity of pleasure they generate is the same, because poetry generates pleasure of higher *quality*.
        2. But what does this mean, and what exactly is the rule for when one pleasure is better than another?
        3. Sometimes, Mill seems to suggest that a 'higher' pleasure is to be *lexically* preferred to a 'lower' one – meaning that it is *always* better to have more of the higher pleasure, *regardless* of how great the cost in sacrifices of lower pleasure is.
            a. 'Lexical' orderings: as in a dictionary, where, if the first letter of one word is earlier in the alphabet than the first letter of another word, you don't even need to look at the words' second letters: the first letter is *lexically more important* than the second, in determining the order in which the words appear in the dictionary. The second letter becomes relevant only if the two words are tied as regards their first letters.
            b. On this interpretation, if reading poetry is a higher pleasure than drinking beer, it would be worth giving up beer for

one's whole life, in order to get to read poetry for even one extra second.
        i. Surely that is absurd??
    c. Crisp (in his 'Routledge philosophy guidebook to Mill on utilitarianism') interprets Mill as making this claim.
4. Other times Mill makes a more modest claim: merely that a given quantity of a higher pleasure is *much more* valuable than the *same* quantity of a lower pleasure.
    a. The modest claim is much more plausible.
        i. The pleasure you gain from reading another philosophy article might be more valuable than going out for a beer once, but it's not so much more valuable that it's worth foregoing an entire lifetime's worth of pub visits for it.
    b. But even the modest version will be contested!
        i. Would you really prefer reading another article except insofar as you thought that in the long run, that would lead to more total pleasure?
        ii. And this modest version opens hedonism up to a powerful objection that the lexical version of the view avoids...
5. ...The problem of Haydn and the oyster
    a. Thought-experiment: Would you prefer the life of an oyster to the 77-year life of Haydn, provided that the oyster life was sufficiently long?
    b. Objection to hedonism: Hedonists have to say 'yes'. But that's crazy.
    c. Possible reply: A version of hedonism that treats higher pleasures as lexically better than lower ones does not have the consequence that the oyster's life is better than Haydn's. (But, as we've seen above, the lexical version is problematic for other reasons.)
    d. Other examples in the same sort of spirit: you have had a bad road accident. Your brain is damaged in such a way that it will pack up in two years if nothing is done. (Those two years would, however, be two years of fairly normal life.) Surgeons ask whether you would like them to operate. The operation will restore to you a normal lifespan. The catch is that it will almost entirely remove your intellectual abilities, leaving you to live the life only of an infant.
        i. Quite aside from issues of being a burden on others, it's not immediately clear which of these outcomes is better *for you*. The issue turns on the relative value of 'higher' and 'lower' pleasures. (Mill would, presumably, always decline the operation - ?)

6. *Both* interpretations are suggested in a single sentence when Mill writes: "If one of the two [pleasures, e.g. poetry and push-pin] is, by those who are competently acquainted with both, placed so far above the other that they prefer it, even though knowing it to be attended with a greater quantity of discontent, and would not resign it for any quantity of the other pleasure which their nature is capable of, we are justified in ascribing to the preferred enjoyment a superiority in quality, so far outweighing quantity as to render it, in comparison, of small account."
    a. The "for any quantity" strongly implies the lexical view.
    b. The "of *small* account" fairly strongly suggests the more modest view (why not say: of *no* account?)
    c. So it's *not clear* how to interpret Mill here.
7. Mill sometimes writes as though the higher/lower distinction coincides with the intellectual/sensual distinction. (This would make the distinction a two-category affair.)
    a. On reflection, this is clearly too simplistic to be plausible. One can distinguish between pleasures of more 'animalistic' and those of more 'refined' type, even *within* each of the categories (intellectual, sensual).
        i. Example within the intellectual sphere: fine appreciation of poetry or philosophy vs. reading a trashy novel
        ii. Example within the sensual sphere: chips and cheese vs a dinner at Le Manoir
    b. One might offer Mill a modification: pleasures are arrayed along a *continuous scale* in terms of quality, rather than simply being "high" or "low"
        i. For our example above: Perhaps: chips and cheese is the lowest, then the trashy novel, then the fine dinner, and poetry appreciation is the highest.
    c. Note that this modification makes the "lexical" view even more implausible than it would be on a "two-category" interpretation.
8. *How can we know* when one pleasure is 'more valuable' than another?
    a. Mill's answer: by seeing what people *who are fully acquainted with both types of pleasure* prefer. "From this verdict of the only competent judges, I apprehend there can be no appeal. On a question which is the best worth having of two pleasures, or which of two modes of existence is the most grateful to the feelings, apart from its moral attributes and from its consequences, the judgment of those who are qualified by knowledge of both, or, if they differ, that of the majority among them, must be admitted as final." (U 2.8)

i. (Note the anti-intuitionist flavour of this remark.)
ii. Mill would give the same reply to the assessment of which of two pleasures/pains of the same 'quality' is more *intense*.

b. Example: Eating chocolate vs reading Tolstoy.
i. Perhaps a competent judge would never give up reading Tolstoy for chocolate, regardless of the (quality or) amount of chocolate involved, or of the timespan over which he would get to enjoy the chocolate.

c. Mill admits that *some* people prefer beer to poetry, fine food to creative pursuits, sex to true love. But he thinks that that only happens when the person in question is not familiar with – perhaps because she is unable to experience – the 'higher' pleasure in question.
i. "It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied. And if the pig, or the fool, are of a different opinion, it is because they only know their side of the question. The other party to the comparison knows both sides." (U 2.6)

d. Objection: It's just not true that 'competent judges' always prefer a higher to a lower pleasure. Even the most dedicated philosophy student *sometimes* prefers going for a beer to reading a philosophy article.
i. Mill's reply: This might happen because the student has already read so much philosophy that she has become temporarily incapable of gaining much pleasure from yet another article. Otherwise, while the student might *choose* the beer over the article, it's not clear that she really *judges it to be more valuable*. She might be choosing it merely from "temptation" and "infirmity of character".
1. This sort of thing happens all the time: cf. failed attempts to lose weight.
2. "It may be questioned whether any one who has remained equally susceptible to both classes of pleasures, ever knowingly and calmly preferred the lower." (U 2.7)
3. <u>Exercise 3</u>: *is* this always what's going on when you choose one of Mill's "lower" pleasures over a "higher" one?

9. Problems with the "test by appeal to the competent judges"
a. It's not clear that there are really any competent judges. The intelligent person may know what it's like *for him* to eat

chips and cheese, but he doesn't know what it's like to live the life of the 'fool' (or to experience eating chips and cheese as the 'fool' experiences it).

b. Is it supposed to apply to thought-experiments involving foregoing a pleasure of one ("higher") type *altogether* in order to increase access to pleasures of another ("lower") type, or are the relevant thought-experiments instead ones involving forgoing a "higher" pleasure *on one occasion* in order to increase access to a "lower" pleasure?

i. It's very plausible that the two questions would receive different answers. I wouldn't give up reading Tolstoy altogether, i.e. choose a life in which I never read Tolstoy, just for more chocolate. But I might well skip one Tolstoy novel, or one instance of re-reading Tolstoy, for more access to (esp. fine) chocolate.

iv. Another 'objection' to Mill on higher and lower pleasures: This aspect of Mill's doctrine amounts to abandoning hedonism - the view that only pleasure matters.

1. Sidgwick: "if... what we are seeking is pleasure as such, and pleasure alone, we must evidently prefer the more pleasant pleasure to the less pleasant: no other choice seems reasonable, unless we are aiming at something besides pleasure." (Sidgwick 1907, pp.94-5)

a. Doubts about this argument:

i. One can be seeking money and money alone (from one's job, say), without necessarily always preferring more money to less. (It might be, e.g., that beyond a certain threshold one doesn't care whether or not one earns *yet more* money.)

ii. Similarly: If I am lonely, I might go to the pub seeking company and company alone, but that does not entail that I care only about the quantity of company (how many people I run into, and how long they spend talking with me?), and not at all about its quality (how well I know and like the people, how interesting the conversations are).

iii. So there is no true *general* principle to the effect of "if you are seeking X alone, then you must always prefer more X to less". Why think that the particular instance of this principle, with "X" replaced by "pleasure", is true? (To put the question another way: What *true* general principle *is* it an instance of?)

2. Moore: "If one pleasure can differ from another in quality, that means, that *a* pleasure is something complex, something composed,

in fact, of pleasure in addition to that which produces pleasure. ... Mill... in admitting that a sensual indulgence can be directly judged to be lower than another pleasure, in which the degree of pleasure involved may be the same, is admitting that other things may be good or bad, quite independently of the pleasure that accompanies them." (Moore, 1903, sec. 48)

a. Doubts about this argument: Mill needn't agree with Moore that "*a* pleasure is... composed... of pleasure in addition to that which produces pleasure", and that differences in "quality of pleasure" just are differences in what produces the pleasure. He could simply hold that the pleasure *itself* has a complex structure, including quality in addition to quantity, and that both aspects of its structure matter.

3. In any case: Moore and Sidgwick, along with many others, have concluded that a higher/lower distinction along the lines of Mill's is inconsistent with hedonism. (Sidgwick rejects the higher/lower distinction; Moore rejects hedonism.) (See also Crisp, pp. 32-4.)
4. A simpler argument: Hedonism *just is* the view that the best life (for the person living it) is the one that contains the greatest net quantity of pleasure. Therefore, in holding that not only quantity but also 'quality' matters, Mill is clearly abandoning hedonism.
    a. But this argument turns the issue into a merely verbal one, turning directly on the stipulative definition of "hedonism". Mill needn't care about merely verbal issues.

e. The master objection to hedonism: The experience machine
    i. Thought-experiment: Would you enter Nozick's 'experience machine'? (Cf. Nozick, Anarchy, State and Utopia, pp.42-5)
    ii. Objection to hedonism: Hedonists have to say yes. But that's crazy (?).
    iii. Hedonists (have to) 'bite this bullet' (i.e., accept a consequence that many people would regard as implausible). But note that many hedonists do so quite cheerfully.
    iv. Related (real-life, non-science-fiction) issues
        1. Is it bad for you if, *unbeknownst to you*, your best friends bitch about you behind your back?
            a. Hedonists have to say 'no', provided that the bitching really is in secret, does not affect how the friends behave to your face, does not affect whether they invite you to events, etc.
            b. Against the hedonist position on this issue: Most people would be very upset even if they found out years later, of friends they no longer had contact with, that this had happened. If the bitching isn't bad for one, then being upset about it seems irrational. But being upset in this case does not seem to be irrational.

Lecture 2: Theories of individual welfare II: Desire-satisfaction and objective-list theories

3. Recap: The experience-machine objection to hedonism
    a. It seems(?) plausible that the life of someone in an experience machine is not going as well *for that person* as an experientially identical non-machine life.
    b. What's missing?
        i. The person in the experience machine might be *having experiences as of* climbing Everest, having wonderful friends and family, wiping out world poverty, etc. But she doesn't actually have any friends, isn't actually in contact with any of her family, and hasn't actually achieved any of the things she thinks she's achieved.
        ii. What she *wanted* was not just to have experiences that were *as if* she was climbing a mountain/having great friends/reforming a country/etc, but *actually to do* those things.
        iii. This suggests that what's missing from the hedonist theory of individual welfare is: Due attention to whether the person's *desires* are really satisfied.
    c. This line of thought motivates the *desire-satisfaction* (or *preference-satisfaction*) theory of individual welfare.
4. Desire-satisfaction theory: What's good for a person is for her desires to be satisfied to the greatest degree possible.
    a. Some examples to illustrate what this means in practice
        i. Our character in the experience machine would have had a good life if she *really had* had lots of close friends, spent lots of time with family, climbed a mountain and eradicated world poverty (or most of those).
        ii. Reading poetry makes Mill's life better, because he *wants* to read poetry. But reading poetry wouldn't make the life of a hamburger-eating bricklayer with no desire to read poetry any better.
        iii. Your life goes better for you if you get a First Class degree if, and then only because, you *want* to get a First Class Degree (or you want to do some things that getting a First will help to cause, e.g. earning ridiculous amounts of money, or pursuing a research career). If you don't happen to want any of those things, then getting a First wouldn't make your life any better.
        iv. Pleasure is good for most of us. But, according to the desire-satisfaction theory, that's just because most of us happen to *want* pleasure.
            1. If there is, somewhere in some monastery, a monk who has managed to transform his psychology to such an extent that he really doesn't care at all about pleasure, then experiencing pleasure would not make *his* life better.
    b. An initial worry about the desire-satisfaction theory: misguided desires
        i. Suppose that I want to get to Cardiff by 9.30pm tonight, in order to have tea with my mother before she goes to bed. Because of this, I have formed a desire to catch the 6 o'clock train out of Oxford. And indeed I do catch that train. But as things turn out, the 6 o'clock train is delayed for two hours by signal failure at Swindon, and doesn't arrive in Cardiff until 10.15. The 7pm train – which I would have caught if I'd missed the 6 o'clock – suffers no such delay, and arrives at 9.15.

ii. It doesn't seem plausible to say that in this scenario, catching the 6 o'clock train *makes my life go better, for me* – it just had the result that I sat on a train for 4 hours and missed my mother entirely.
iii. But I did have a desire to catch the 6pm train! So doesn't the desire-satisfaction theory entail that catching that train is indeed good for me?
iv. Obvious reply: My desire to catch the 6pm train *doesn't count*, for the purposes of determining what's good for me. I only *had* that desire because I thought it would serve another, more fundamental desire I had, viz. the desire to see my mother – and that belief turned out to be mistaken.
v. Clarifying this reply:
    1. The point is that the desire-satisfaction theorist needs a distinction between *instrumental* and *final* desires.
    2. A final desire for X is a desire for X *just for its own sake*. An *instrumental* desire for X is a desire for X that one holds only because one believes that X is a causal means to securing Y, and one has a desire for Y (which latter may be another instrumental desire, or a final desire).
    3. This is very similar to the distinction that the hedonist needed to draw between things that are good in themselves (have 'final value'), and things that are good because they are causal means to other good things (they have 'instrumental value'). (Recall that to be at all plausible, the hedonist needed his claim to be that only pleasure has *final* value, not that only pleasure has value *full stop*.)
vi. Clarified version of desire-satisfaction theory: A person's life goes well for him to the extent that his *final* desires are satisfied.

c. The arithmetic of the desire-satisfaction theory
    i. How do we determine 'the extent to which someone's final desires are satisfied'?
    ii. Attempt #1: *Count* his desires. The goodness of a life is just the number of satisfied desires it contains.
    iii. An objection to the 'just count' version: Surely the *strength* of the desires matters too?
        1. E.g. If I have a very strong desire to climb a mountain, and two very mild desires (say, for a chocolate ice cream and to stroke a cat), surely it can be better for me to have one stronger desire satisfied than to have two weaker desires satisfied.
        2. This suggests an alternative 'arithmetic of desires', something like: Assign to each desire an *intensity*. Add up the intensities of the satisfied desires within the life you are evaluating. The resulting figure is the measure of how good the life is for the person living it.
            a. This is a *bit* like the Benthamite 'hedonic calculus' method of adding up pleasures.
            b. This analogy also suggests: perhaps the length of time you held the desire for matters too: long-lasting desires count for more than short-lived desires.

d. Objections to desire-satisfaction theory: surely not all desires (not even all *final* desires) count?
    i. Other-regarding desires
        1. Very few people are completely selfish. Suppose I have a desire that tropical diseases in the third world are eradicated. It doesn't seem to follow that, if those diseases are indeed eradicated, that makes *my* life better. But according to desire-satisfaction theory, that would follow.
        2. A possible reply: if I'm deeply committed to the project of eradicating poverty – if, say, a large part of my life's work has been dedicated to that cause – then it's not so implausible to say that my life goes better if the goals of that project turn out to be fulfilled.
        3. Rejoinder: But other-regarding desires can occur even where there is no plausible "project" story to tell. E.g. Suppose that I meet a stranger on a train, and, chatting with him, form a strong desire that he should flourish. We then part ways, and I never see or indeed think of him again. Unbeknownst to me, he does flourish. It is not at all plausible to say that this makes *my* life go better.
    ii. Immoral desires
        1. Suppose that Jim is a sadist, and strongly desires that others suffer pain. It does not make Jim's life go better if others do indeed suffer pain (at any rate: if he does not get to witness their pain).
        2. Objection: Our 'not good!!' gut reaction to this case is a result of the tendency of Jim's character to make the lives of *others* go badly, perhaps together with the thought that Jim doesn't *deserve* to have his desires satisfied. It need not be taken as a reliable indication that the fulfilment of Jim's sadistic desires does not contribute to making Jim's life go well *for him.*
e. Modified version of desire-satisfaction theory: a person's life goes well for him to the extent to which his self-regarding preferences, i.e. preferences regarding his own life, are satisfied. (Parfit calls this the 'success theory'.)
    i. A preliminary problem with this theory: Which desires count as being 'about my own life'? Some unclear cases:
        1. Desires for the success of my projects:
            a. If I have spent much of my life working to eradicate poverty, does it make my life go better if my efforts succeed, even if the success is unknown to me e.g. because it postdates my death?
            b. Parfit's example: If I have a strong desire *to be a successful parent* and if one of my children's lives goes badly as a result of my parenting mistakes, that makes my life worse even *for me* (not just: for my child). (Hence Parfit would also answer 'yes' to the question in (a) above.)
                i. In contrast, Parfit *doesn't* think that if one of my children is killed in an avalanche, but I never find

out, that that makes my life go worse for me – even if one of my strongest desires is *that my children flourish*.

ii. If that's right, then the connection to the agent's own projects – to things that she hoped to do or achieve – is crucial, in drawing the line between desires whose satisfaction does improve one's life and desires whose satisfaction does not.

ii. But anyway, there seem to be some preferences that are uncontroversially 'about my own life', but whose satisfaction still doesn't make my life go better. If that's so, then success theory cannot be correct either.

1. Irrational desires
    a. Basic thought: Some desires are *irrational*, and it doesn't make one's life go better to have one's irrational desires satisfied (instead, one's life would be made better just by getting rid of the irrational desire).
    b. Example 1: the drug addict
        i. Suppose an addict has a strong desire for a heroin shot. But suppose also that getting the shot, while it brings temporary relief and some transient pleasure, would only fuel his addiction, making future episodes of unfulfilled craving more extreme, and reducing his chances of getting off the drug and getting his life back on the rails. It seems distinctly odd (doesn't it??) to say that getting the shot *makes his life go better for him*, just because it satisfied his desire (this particular 'desire' might more naturally be called: a craving).
        ii. A possible response to the drug-addict case: the drug addict has desires that *conflict* with one another. He wants the immediate fix, but he also wants to break his habit and return to normal life. The reason we wouldn't say that getting the fix makes his life go better overall is that while it satisfies one desire, it frustrates another, and probably stronger, desire.
        iii. Reply: Not all drug addicts *do* have that second desire. Even in the case of a drug addict who lacks the desire to get better, it doesn't seem right to say that getting the heroin shot makes his life go better.
    c. Example 2: desires for ever-expanding material wealth
        i. Most people desire more money, and more of the things money can buy: bigger houses, fancier electronic goods, more expensive clothes, more meals out etc.

ii. But research (both formal and anecdotal) arguably suggests that above a certain (and quite low!) threshold, increasing material wealth *doesn't actually make people any happier*.
    1. Having less wealth than other people whose lives you interact with and/or observe often makes one unhappy. But the absolute level of wealth itself seems(?) to be rather unimportant.
    2. (Note: many economists (among others) would strenuously deny this. Decide where *you* stand on this issue!)
iii. Desire-satisfaction theory entails that increasing someone's material wealth (provided she wanted that increase) makes her life go better for her, even if it fails to make her any happier. Many people would, at least with the benefit of hindsight, deny that, even in their own cases.

d. So, even if the desire-satisfaction theorist accepts the restriction to self-regarding desires, her theory still seems to get some of the crucial cases wrong.

f. Unified response to these objections: the desires that count, for the purposes of desire-satisfaction theory, are just the desires for things that *are in fact good for us*.
    i. But that means that the desire-satisfaction 'theory' needs to be supplemented by an independent account of what is good for a person. And at that point, the appeal to desire-satisfaction is no longer seems to be doing any work.

g. Comment on the dominance of the preference theory in economics
    i. Something like the desire-satisfaction theory appears to dominate much of the normative (as opposed to descriptive) theorising that goes on in economics.
    ii. Economists construct 'utility functions' according to which one state of affairs has higher 'utility' for a given person than another if and only if the person *prefers* the first to the second. (We might be comparing, e.g., states of affairs in which one works for N hours per week, has 168-N hours of 'leisure' time and consumes M units of goods and services, for variable values of M and N.)
        1. Note that this is *not* 'utility' in the *utilitarian* sense of that term. Utilitarian utility is about *happiness*, not about preference-satisfaction.
    iii. These utility functions are sometimes used just to make *predictions* about what people will in fact do. ('Descriptive economics.') For that purpose, people's actual preferences are clearly the right thing to use: what people will in fact do is determined by which preferences they actually have, and

issues of whether those preferences are themselves irrational or misguided are irrelevant.

iv. But other parts of economics are *normative* – they are concerned with what governments (in particular) *ought* to do. The use of 'utility'-maximising principles in normative economics, if by 'utility' is meant a quantity that represents simply the degree to which a person's preferences are satisfied, amounts to buying into a desire-satisfaction theory of well-being.
    1. An example of such a principle: the Pareto principle

v. 'Gross national product' vs 'gross national happiness'
    1. The government of Bhutan seeks to maximise GNH, not GNP. This appears to be an operationalisation of an hedonistic, as opposed to a desire-satisfaction, account of well-being.

h. Another worry about the desire-satisfaction theory of individual well-being: Doesn't that theory anyway 'get things backwards'?
    i. Suppose you are having the sort of 'early life crisis' that many of us experience around-about our early twenties, when we're trying to work out (as we'd normally put it) what's important, what our values are. You're not sure, for instance, whether the good life consists in riches and luxury (material consumption), achievement, or spiritual enlightenment, and hence you don't know, for instance, whether to aim for a career in banking, research or the monastery. Your desire-satisfaction-theorist friend tries to reassure you. "The good life is simply getting what you happen to want. So it's easy – just introspect to find out what your desires are, and that'll be the answer to what's good for you." This answer seems to miss the point: phenomenologically, it seems that what's going on is rather: you're trying to work out on some independent ground what's important, and *once you have*, your desires will then change to track that discovery.
    ii. A similar objection could be raised against hedonism. One very important source of pleasure is living a life that you believe is a good one…

i. On the other hand, the "experience machine" thought-experiment did seem to provide a good reason for doubting hedonism.

j. This raises the question: Is there any other theory of individual welfare that
    i. would support the conclusion that a life in the experience machine is not the best life for the individual, but also
    ii. doesn't suffer from the problems that desire-satisfaction theory suffers from either?

k. *Objective list theories* are supposed to fit the bill...

5. Objective list theory
    a. The basic idea of objective list theories: Some things (e.g. knowledge, friendship, achievement appreciation of beauty) are just good for people, regardless of whether people desire them and regardless of whether they are causes of pleasure. Other things (e.g. excessive material consumption, harmful drugs) are not good for people, even if people desire them.
        i. The "objective list" is the list of items that are objectively good for people.

ii. On any plausible version of objective list theory, pleasure/happiness will be included on the list.
    1. Thus objective list theory retains some of the key insights of hedonism.
    2. Hedonism is in fact a special case of objective list theory: hedonists are those who insist that the list contains only one item, namely happiness/pleasure.
iii. Other candidates for inclusion on the list:
    1. Knowledge/understanding
    2. Friendship/other personal relationships
    3. Achievement
    4. Aesthetic appreciation
    5. Autonomy
iv. According to objective list theory, the life in the experience machine is not the best life for the individual because, although it contains the most possible pleasure, it is severely deficient in terms of most or all of the other items on the list.

b. Moore's version of objective list theory (*Principia Ethica*, ch. 6; see http://plato.stanford.edu/entries/moore-moral/#3 for a helpful summary)
    i. Caveat: Moore is actually talking about which things are good *overall*, rather than (our current question) what is good-for-particular-people. (Moore's own view is that the notion of "good for you"/"good for me" makes no sense, unless it just means: good, and in addition, contained in your life/in my life. In Lecture 1, I argued against that: Recall the example of the man who inadvertently discovered a cure for cancer, but led a miserable life.) But most of his discussion would go through equally well if it was about what was good *for* individual people.
    ii. Moore's methodological test for estimating an object's intrinsic value: how much value would that object have if it existed *all on its own*?
        1. Example of the application of this method: Imagine first that no sentient creatures exist, but that the world is very beautiful. Is that state of affairs better than, or of equal value to, one in which no sentient creatures exist and the world is ugly?
            a. Moore's judgment: The beautiful world is better.
            b. Conclusion to draw from this: beauty has positive intrinsic value, even when no sentient creature is aware of it.
        2. One point of this method is to prevent judgments of how valuable a thing is *itself* (what its 'final value' is) from being infected by knowledge of which other things it tends to *bring about*. (The values of the things it brings about contribute to its *instrumental* value, but not to its *final* value.)
            a. In the above example, the test enables us to conclude that beauty has *final* value, not merely instrumental value.
            b. In contrast, Moore thinks that the same test shows that pleasure on its own has no final value: he thinks that

pleasure has value only insofar as it leads to e.g. *awareness* of pleasure.

iii. Moore's "principle of organic unities": the value of a "whole" that is made up of several parts need not be equal to the value of the sum of the parts.
    1. Moore's view on punishment: Vice (a bad state of mind) is intrinsically bad. And pain is intrinsically bad. But the state of affairs in which pain is inflicted on a vicious person is *not as bad as* a simple summing of the 'badnesses' of vice and of pain would lead us to expect.
    2. Even if (contra Moore's own view) there *were* no intrinsic value in a world that contained no pleasure (but did contain e.g. knowledge and contemplation of beautiful objects), it would not follow that pleasure is the only valuable thing, still less that the value of a state of affairs is represented simply by the *sum* of the values of all the pleasures it contains. Instead, it could well be that: knowledge and contemplation of beauty are of no value *if unaccompanied by pleasure*, but when they *are* accompanied by pleasure, the value of the resulting whole is much greater than the value that the pleasure would have if it existed all by itself (i.e. not as a response to knowledge/beauty).

iv. On the contents of Moore's list (in *Principia Ethica* – not necessarily in his later work!)
    1. Moore takes the most valuable things to be cases of aesthetic appreciation of inanimate objects, and cases of 'personal affection'.
    2. In both cases, the valuable thing is a *complex whole*, consisting of an object contemplated (e.g. a painting, an opera, another person), a 'cognition' (perception) of that object, *and* an emotional reaction to that cognition that is appropriate to the object's qualities.
    3. Examples in the aesthetic case:
        a. An opera, a hearing of that opera, and a profound sense of tragedy on contemplating the fate of the opera's doomed-lover protagonists.
        b. A natural landscape, a sighting of that natural landscape, and an appreciation of the grandeur of nature.
    4. Examples in the personal-affection case:
        a. A person with fine aesthetic taste, a second person's perception that the first person does have that good taste, and the second person's admiring the first on that account.
        b. A courageous and compassionate person, and the analogous two additions.
    5. Moore takes many things that would often be accorded prime position on an "objective list" to have "little or no intrinsic value" in themselves, although they might contribute to the great value of certain complex wholes of which they are parts:
        a. Pleasure

        b. Knowledge
    6. Despite thinking that pleasure has very little (positive) intrinsic value, Moore does think that *pain* has a large amount of *negative* intrinsic value: he thinks that there is an asymmetry between pleasure and pain.

v. Criticisms of Moore's list
    1. In emphasising the value of passive contemplation, Moore fails to recognise the value of more active forms of interaction with the world. Examples:
        a. *Achievement* is not on Moore's list at all.
        b. Moore's account of personal affection is peculiar. It involves only the "admiring contemplation" of the other person's fine qualities, not e.g. any desire for two-way interaction with the other person (helping him in his projects, conversing with him, having any physical or sexual relationship with him).
    2. Moore under-rates pleasure. (Relatedly: his claim of a pleasure-pain asymmetry is implausible.)
    3. Some features of the list just seem arbitrary: e.g. it's weird to *include* beauty, but *exclude* knowledge.

vi. Reply (endorsed by Moore himself): Moore's discussion is more useful for its account of *what the key questions are* and of the *method for answering them*, than for the particular answer it suggests.
    1. Indeed, Moore himself changed his mind later about several of his claims regarding what is, and is not, on the list.

c. The paternalism/arrogance objection to objective list theories
    i. Consider a person who neither wants nor gains any pleasure from some item on the objective list theorist's List.
        1. E.g. a hermit with no desire for friendship; an ignoramus with no interest in learning.
    ii. The objective list theorist appears to be committed to saying that we would make this person's life better *for her* if we induced her, against her will, to form friendships/acquire knowledge/etc.
    iii. Objection:
        1. This conclusion is objectionably *paternalistic*. We have no right to interfere with the person's life, against her will, in this way.
        2. The objective list theorist is intolerably *arrogant*, in assuming that his theory of the good is correct even for the person whose own implicit theory of the good disagrees with that list.

d. Reply to the arrogance objection
    i. This objection relies on an implicit premise that disagreeing with someone implies disrespecting them/behaving arrogantly towards them.
    ii. But we don't normally accept that principle. Some examples outside the context of goodness/morality:

1. I believe that the Earth is (approximately) round. There exist people who think that it is flat. I disagree with them: I think that their beliefs are *false*. This does not imply that I disrespect them, nor that I am behaving arrogantly towards them.
2. I believe that there is no God. There exist (many!) people who believe that there is a God. I think that their religious beliefs are false, but it doesn't follow from that that I disrespect them or that I am arrogant.
3. A scientific example: I believe that the best interpretation of quantum mechanics is a many-universes theory. Others disagree. I think that certain of their scientific beliefs are false, but I have the highest respect for them.
4. Note that in no such case do I need to agree that the Earth is flat "for" the other person, that there is a God "for" him or that a many-universes theory is false "for" him, in order to be according him due respect (unless that talk of "for him" is just a confusing way of reporting what his beliefs are).

iii. Similarly: The believer in objective list theory is committed to thinking that the hermit/the ignoramus is missing out on some things that would be good for him, but it doesn't follow from this that there is any disrespect or arrogance going on.

e. Reply to the paternalism objection
   i. The claim that e.g. friendship would be good for Joe does not entail the claim that it would be right for a third party to intervene in Joe's life, against Joe's will, in order to cause Joe to have friendships.
   ii. Two reasons why this entailment does not hold
      1. On any plausible objective list theory, *autonomy* is likely to be one of the key items on the list.
         a. A person's life is better for him, other things being equal, if he is the author of that life, choosing how to live it in accordance with his own values.
         b. If some of his values are mistaken, then there is a tradeoff between the sacrifice of autonomy that would be involved in intervening in the course of his life on the one hand, and the gain along the other dimensions of well-being that such intervention might effect.
         c. In many cases, the sacrifice of autonomy would be so bad that the 'trade' is not worthwhile: thus, despite the fact that as things stand Joe e.g. has no friends, we would still make his life worse overall if we intervened to force a more sociable lifestyle on him.
      2. In any case, claims about what would make someone's life *better* do not immediately entail any claims about what one *ought to do* to that person:

a. "What would make someone's life better": this is about what is <u>good</u> for him.
b. "How we should treat him": this is about what we <u>ought to do</u>/which ways of treating him would be <u>right/wrong</u>.
c. Questions of which outcomes would be good on the one hand, and questions of what's right to do on the other, are distinct questions.
d. A non-consequentialist (for example) might well think that even if intervention would make someone's life go better, still it would be wrong to intervene, because the person <u>has a right</u> to live his life the way he chooses, and we <u>have no right</u> to intervene.

f. A more serious objection to objective list theory
   i. Consider the hermit again. Suppose not only that he currently has no desire for friendship, but also that even if we did intervene to change his lifestyle and brought it about that he had several genuinely close friendships, he would still take no pleasure in those friendships, and would spend his days thinking wistfully back to the hermit lifestyle he was wrested from.
   ii. Or consider the ignoramus. Suppose that we get her through GCSEs and send her on to university; suppose further that, completely intellectually reformed, she does quite well at university, and develops quite a deep understanding of her subject. But suppose that throughout, she takes no pleasure in any of this, and that, contra Mill, she really would choose to have remained ignorant if she could turn the clock back.
   iii. Intuitively, it just doesn't seem plausible to insist that the lives of these two characters have been improved by the addition of the "missing" goods from the objective list, given that they *are not enjoying* and *do not want* possession of those goods. Here, the hedonist and desire-satisfaction theories that we have (tentatively) rejected seem to have captured an important insight that the objective-list theory lacks.

6. Hybrid theories (Some suggestions of this in Moore; see also Parfit, *Reasons and Persons*, pp.501-2; Kagan, 'Well-being as enjoying the good', *Philosophical Perspectives* 23, 2009)
   a. This last observation provides the motivation for *hybrid* theories of well-being: theories that attempt to combine key elements of two or more of the above theories, in such a way as to capture the successes of each theory while avoiding the objections to any non-hybrid theory.
   b. One hybrid suggestion is: well-being is made up of two components, both of which need to be present in order for the person to be well off. First, one must possess something that is objectively good. Second, one must enjoy possessing that thing. On this theory:
      i. The person in the experience machine is not well-off because she does not *possess* most of the objective goods.
      ii. The manipulated hermit is not well-off because, although he *possesses* many objective goods, he is not *enjoying* them.

c. Kagan's paper helpfully surveys some of the considerations that one will need to deal with in order to develop a fully worked-out version of this sort of theory.

Lecture 3: The theory of the good and the criterion of right action

7. Structural comments:
    a. Many people (not all!) see ethics as having a three-part structure:
        i. Theory of individual welfare: What makes a person's life go well *for her*? (Discussed in the previous two lectures.)
        ii. Theory of the (overall) good: Given an answer to the first question, what makes one state of affairs better than another *overall*?
        iii. Criterion of right action: Given an answer to the first and second questions, what does it take for an action to be right/wrong?
    b. This lecture tackles the second, and starts to tackle the third, of these.
8. Overall goodness
    a. It's natural to think that increasing the extent to which people's lives go well for them is at least *part* of improving the state of affairs overall.
    b. *Welfarism* is the (more controversial) thesis that the overall goodness of a state of affairs is *entirely determined* by how well the lives lived in it go for the people who live them.
    c. Some reasons to doubt welfarism
        i. One might hold a view of well-being according to which living one's life *autonomously* is no part of what makes one's life go well for oneself – that is, that autonomy is no part of well-being – but nevertheless think that a state of affairs in which persons live their lives autonomously is much better than a state of affairs in which persons are entirely controlled by outside forces, but have the same welfare levels.
            1. This is probably not a very convincing objection, since if you really did value autonomy in any such way, you probably *would* have included it in your account of well-being in the first place.
        ii. One might hold that the mere existence of certain things – for instance, great works of art, or well-functioning intact ecosystems – is good in itself, quite independently of any contribution those things make to the well-being of any human or animal.
    d. There are several varieties of welfarism (since to say *that* A determines B is not to say *how* A determines B).
        i. Utilitarianism: The overall goodness of a state of affairs is just the *sum* of people's well-being levels.
    e. A common objection to the utilitarian theory of the good is that, in paying attention only to *total* welfare, it ignores the importance of the *distribution* of welfare amongst persons.
    f. Some alternatives:
        i. Prioritarianism: The contribution that an extra unit of well-being makes to overall goodness depends on the existing well-being level of the person it accrues to: its contribution is greater if it accrues to a less well off person. ("Well-being has diminishing marginal overall value.")

ii. (Welfare-)Egalitarianism: Other things being equal, a state of affairs is better if it involves a *more equal* distribution of well-being among persons. ("Equality is intrinsically important.")
iii. The distinction between prioritarianism and egalitarianism is a bit subtle. In the final analysis, it turns out to be a *technical* issue, and one that will not concern us in this course.

g. However, even if the utilitarian's theory of overall good was replaced by a prioritarian or egalitarian one, most of the most interesting objections to the theory would still remain, since they are objections either to hedonism or to the utilitarian's criterion of right action.

9. Criterion of right action
   a. To say *which state of affairs is best* is not immediately to say anything about *what anyone ought to do*.
   b. Maximising act-consequentialism
      i. Criterion of right action: An act X is right iff (that is: if and only if), of all acts that were available to the agent at the time of action, X leads to the best consequences; otherwise it is wrong.
      ii. Simple examples
         1. In the classic 'trolley problems' (see, e.g., Thomson, "Killing, letting die, and the trolley problem"), maximising act-consequentialism always holds that killing the one (or letting the one die) in order to save the five is right, and that the alternative course of action would be wrong.
         2. In 'white-lie' cases, maximising act-consequentialism might recommend either telling the lie or telling the truth, depending on how it ranks the goodness of knowledge vs avoidance of discomfort in its theory of the good.
            a. Contrast with this maximising act-consequentialism:
               i. an 'absolute deontological' theory according to which it is simply wrong to lie, regardless of the consequences;
               ii. a non-absolute deontological theory, according to which the wrongness of lying is such that sometimes one should not lie even though the consequences would be *a bit* better if one did.
      iii. Some more complex examples
         1. On vegetarianism:
            a. Act-consequentialism would be concerned with what the consequences of one's eating this particular piece of meat would be on (i) the number of animals who are killed, (ii) the number of animals who are born, (iii) the welfare of those animals who do live, during their lives.
            b. It would therefore probably recommend eating some types of meat but not others, and would recommend eating meat in some circumstances but not others.

2. On abortion:
    a. Act-consequentialism would weigh up (i) the goodness of the life that the foetus would go on to have if it lived, for the person that the foetus would become, (ii) the extent to which that life would benefit or harm others (for example, if the foetus goes on to discover a cure for cancer, or commits mass murder) in general, (iii) the effect on the mother's quality of life in particular, (iv) the effects on how many other children the mother goes on to have in the future, (v) issues arising from the fact that the foetus, if allowed to life, would *itself* be quite likely eventually to have children.
    b. It would therefore probably permit some abortions but not others, depending on how factors (i)—(iv) pan out in the individual case in question.
3. Obviously, these issues are complex. But arguably, that's not act-consequentialism's fault – it's no virtue of a theory to take difficult, complex questions, and pretend that they're simple.

c. Objections to maximising act-consequentialism
    i. Cluelessness objection:
        1. The objection: One never knows, at the time of action, what all the consequences of one's action will be. Therefore one never knows, at the time of making one's decision, whether a given action will be right or wrong according to this theory. So:
            a. The act-consequentialist criterion is useless as *guidance* to the decision-making agent; and
            b. It's inappropriate to *blame* someone for doing something that is 'wrong' in the act-consequentialist sense.
        2. Examples:
            a. The case of abortion, sketched above.
            b. More general point: All of our actions have enormous numbers of possible long-run consequences that we can't possibly predict.
                i. Analogy: a butterfly flapping its wings in Texas can cause a storm in Bangladesh
                ii. Human-action cases: Stopping to let a stranger past might cause her later to be run over by a bus, or to fail to get the job she is being interviewed for, or not to meet the person who would have been her life partner...
        3. Reply, part 1: All this is correct, but it doesn't mean that the act-consequentialist criterion has no place in moral theory.
            a. The criterion does not itself say anything about *how one should make one's decisions*, or *when to blame someone*. These could be the tasks of some part of moral theory *other than* the criterion of right action.

i. One might well want to say, e.g. of a white lie that was unfortunately seen through, "telling that lie turned out to be the wrong thing to do (although James can't be blamed for telling it, since he couldn't have known that at the time)."

b. Reasonable principles linking wrongness to blame might be, e.g.:

i. If someone does something *while knowing that* it is wrong, then he is blameworthy for having performed that action.

ii. If someone does something having made *no effort to find out* whether it is wrong or not, then he is blameworthy for being morally reckless.

c. (Question: Does a criterion of right action that is subject to the 'cluelessness' objection have any important place in moral theory? If so, what exactly is its place?)

4. Reply, part 2: We can distinguish between *objective* and *subjective* criteria of right action.

a. An act is *right in the objective sense* iff it *in fact* leads to the best overall consequences; otherwise it is *wrong in the objective sense*.

b. An act is *right in the subjective sense* iff *the agent believed* that it would lead to the best overall consequences;[1] otherwise it is *wrong in the subjective sense*.

c. Suggestion: An agent is *blameworthy* if he performs an act that is *wrong in the subjective sense*.

ii. The demandingness objection

1. The canonical example:

a. I can save a child's life by giving £1400 to a (carefully chosen) charity. According to some estimates, I can keep a child in primary school for an extra year by giving as little as £2. (Source: www.givingwhatwecan.org)

b. Clearly, this money will do more good if I give it to the charities concerned than if I spend it on myself.

i. The basic reason for this is what economists call the *diminishing marginal utility* of money.

c. Suppose, then, that I decide to give £1000 a year to these charities.

d. It will *still be true* that I would do much more good by giving another £1 to charity than by spending it on myself...

e. Until I am as poor as the world's poorest.

[1] Better: if its *expected moral value*, relative to the credences that the agent held at the time of the decision, is highest. 'Expected moral value' here is a probability-weighted sum of the goodnesses of the possible resulting states of affairs (cf. 'expected utility theory' in economics).

f. So maximising act-consequentialism (whether subjective or objective) requires me to give virtually all my money to charity??
    i. And also to spend all my spare time earning as much money as I can in order to give more, at the expense of spending any time with family and friends, having any hobbies, just relaxing, etc.
g. This seems *too demanding*. Doesn't it?
    i. If morality does not in fact demand this much, then both the objective and the subjective act-consequentialist criteria of rightness must be incorrect.

2. Possible replies:
    a. Accept the objection – change the criterion of rightness.
    b. Argue that act-consequentialism does not in fact require this much of agents.
    c. Accept that morality really is this demanding, and that virtually everything we currently do is morally wrong.

iii. The objection from deontological side-constraints

1. Example 1: The Sheriff
    a. Suppose that a sheriff in a small town in South America is faced with a spate of violent crimes. The majority of the townspeople believe that Joe Bloggs, who happens to be in the sheriff's cells at this moment on a minor public disorder charge, is the culprit. The sheriff knows that Bloggs is not the culprit. But he also knows that he has no hope of catching the real culprit, and that unless he publicly hangs *someone* for the violent crimes, the townspeople will riot – and that the riots will lead to several deaths of innocent people. He knows that if he does hang Bloggs, Bloggs' innocence will never be discovered, so that there will be no adverse consequences in terms of e.g. reduced respect for the institutions of law and order. Should the Sheriff hang Joe Bloggs, or not?
2. Example 2: Organ harvesting
    a. Suppose that a doctor has five patients, each of whom urgently needs the transplant of a (different) organ. If they don't get these transplants, they will all die within a few days, but the prospects for finding tissue matches on that timescale are extremely slim. That day, however, a healthy patient comes in for a blood test, and the doctor happens to notice that this patient is a perfect tissue match for all five of the critically ill patients. Suppose that there are no dangers associated with the transplant operations, so that the doctor knows that if he were to kidnap and kill this one

patient, he would certainly be able to save the lives of the other five. Suppose further that nobody would ever discover the kidnap, and that the doctor knows this. Should he abduct and kill the healthy patient?

3. Objection: According to act-consequentialism, the sheriff is not only morally *permitted*, but is morally *required*, to hang the innocent Joe Bloggs. Similarly, the doctor is morally required to harvest organs from the innocent blood-test patient. But in fact these agents are not even *permitted* to do these actions (they would be *wrong*). Therefore act-consequentialism's criterion of right action is false.
4. A possible avenue of reply: in *realistic versions* of these cases, the consequences of killing the One would not in fact be best overall, once all long-run effects are taken into account.
    a. There's a high chance that the word would get out that sheriffs/doctors (resp.) behave this way. And the consequences of *that* would be
        i. A decreased respect for the forces of law and order;
        ii. People would avoid going into hospitals, whether as patients or as visitors, except in situations of direst need.
    b. This sort of line of thought *might* show that *in practice* maximising act-consequentialism does not require the problematic acts. But the act-consequentialist still has to agree that in 'pure' versions of these cases (i.e., in which the sheriff/doctor can be *absolutely sure* that no-one else will hear of what he has done, and that his own character will not be adversely affected by performing the act in question), *then* the theory's implications are as the objector claims.
        i. Question: How bad is that?

iv. The integrity objection

1. Example: George's job choice
    a. "George, who has just taken his PhD in chemistry, finds it extremely difficult to get a job. He is not very robust in health, which cuts down the number of jobs he might be able to do satisfactorily. His wife has to go out to work to keep them, which itself causes a great deal of strain, since they have small children and there are severe problems about looking after them. The results of all of this, especially on the children, are damaging. An older chemist, who knows about this situation, says that he can get George a decently paid job in a certain laboratory, which pursues research into chemical and biological warfare. George says that he cannot accept this, since he is opposed to chemical and biological warfare. The older man replies that he is not too keen on it

himself, but after all George's refusal is not going to make the job or the laboratory go away; what is more, he happens to know that if George refuses the job, it will certainly go to a contemporary of George's who is not inhibited by any such scruples and is likely if appointed to push along the research with greater zeal than George would. Indeed, it is not merely concern for George and his family, but (to speak frankly and in confidence) some alarm about this other man's excess of zeal, which has led the older man to offer to use his influence to get George the job... George's wife, to whom he is deeply attached, has views (the details of which need not concern us) from which it follows that at least there is nothing particularly wrong with research into CBW. What should he do?" (Williams, 'A critique of utilitarianism', in Smart and Williams, Utilitarianism: For & Against, pp.97-8)

2. Act-consequentialism's verdict on this case: George ought to take the job, simply because the consequences of him taking it are better than the consequences of him not taking it; end of story.
3. Williams' objection to utilitarianism's account of such cases: "the integrity objection"
    a. Every agent has a number of (first-order) *projects*.
        i. These include: desires for basics like food, shelter, physical security, for oneself and for one's family and friends; desires for 'objects of taste' (furnishings, artwork); an interest in poetry; the pursuit of philosophy; support of some cause, e.g. the environment, pacifism, economic equality.
    b. On any plausible account of well-being, success in one's (valuable/reasonable) *projects* is a key element of well-being.
    c. So, in particular, the utilitarian has to agree that the existence of first-order projects is important.
    d. The agent obeying utilitarianism must always be guided (though) by the *second*-order project of maximising utility.
    e. This will *sometimes* involve acting so as to pursue his own projects. But utilitarianism will only recommend acting to pursue one's own projects when the causal structure of the situation "just happens" to be such that one can generate more utility that way than by acting to further others' projects (or non-project sources of utility) instead.
    f. Requiring this degree of distance between an agent and his own first-order projects "is to alienate him... from his actions and the source of his action in his own convictions...

It is thus, in the most literal sense, an attack on his integrity."

4. Reply to the integrity objection: the utilitarian agent isn't acting against (or in a manner unrelated to) his own deepest commitments *if his deepest commitment is to utilitarian moral theory*. (Relatedly: Any other moral theory will face the same issue. Requiring an agent *who does not believe in a given deontological ethical theory* to conform to that theory will equally "alienate that agent from the source of his action in his own convictions", but surely the more important question is what happens to agents who *do* believe in the theory in question.)

d. Other forms of act-consequentialism
   i. *Satisficing* act-consequentialism: There is some *threshold* level of overall goodness, such that one's act is right provided its consequences are at least as good as that threshold, and wrong otherwise.
      1. On the demandingness objection: A 'satisficing' moral theory is less demanding than maximising theory (potentially *much* less demanding).
      2. On deontological side-constraints: This objection applies equally to a satisficing theory, since the satisficing theory agrees (with the maximising theory) that one is always at least *permitted* to being about the best state of affairs.
   ii. *Scalar* consequentialism: This theory does not provide a criterion of *right* action. It simply says that one action is *better* than another if its consequences are (overall) better.
      1. On the demandingness objection: Scalar consequentialism simply notes that it is better to give £10 to charity than nothing at all, better again to give £1000, and best of all to give almost all one's money. It is immune to the 'demandingness objection' since it does not make any *demands*.
      2. On side-constraints: Again, scalar consequentialism is equally vulnerable to this objection, since it agrees that (e.g.) hanging the innocent man is the *best* thing the sheriff can do.

Lecture 4: Rule-consequentialism; Mill's account; deontological theories

10. Rule-consequentialism introduced
    a. The consequentialist might attempt to capture intuitive verdicts regarding 'deontological constraints' by focussing not on *individual acts*, but rather on *rules*.
    b. 'Rules': These could be the constraint-theorist's principles 'don't kill', 'don't lie' etc.
    c. An obvious question: Why does morality require keeping to these rules, rather than any old alternative rule ("don't wear purple at latitudes whose decimal expression, rounded to one decimal point, ends in a '3'"?).
    d. It's natural to think that the justification of rules has to have *something* to do with making things better, i.e., something to do with consequences. But there are several different ways in which this could be made more precise...

11. "Compliance" versions of rule-consequentialism: individual vs society, full vs partial compliance
    a. Talk of the 'consequences of the rule' is ambiguous. (The consequences of writing the rule on the blackboard?) *Compliance* rule-consequentialism focuses on the consequences of *complying with* the rule.
    b. But this still leaves open the questions of *whose* compliance with the rule is relevant. We need to distinguish between "individualistic" vs "society" versions of the theory and, within the latter version, between "full compliance" and "partial compliance" sub-versions:
        i. Individualistic compliance rule-consequentialism: The right act is the act that conforms to a rule such that the consequences of *the agent's* complying with that rule are better than the consequences of *that agent*'s complying with any alternative rule.
            1. Example: Lying. The consequences of my *generally* complying with a 'rule' that says "lie whenever it appears that doing so would lead to better consequences" would be that no-one trusts anything I say any more.
            2. But in other cases it seems to deliver the (intuitively!) wrong answers. E.g. *my* complying with the rule "vote in political elections whenever I am eligible" may (?) not have better consequences than my accepting the rule "don't bother to vote", since it is so unlikely that a single vote will make the difference, and given the time and effort involved in voting.
            3. The 'voting' example is a case in which the version of rule-utilitarianism currently under discussion seems too *permissive*. In other cases it seems too *demanding*: e.g. my accepting the rule 'donate 98% of my income to the best charities' seems (?) to have better consequences than my accepting any less demanding rule on charitable giving.
                a. Thus, our first version of rule-consequentialism is just as vulnerable (or not!) to the "too demanding" objection as act-consequentialism is.
        ii. Society-wide full-compliance rule-consequentialism: The right act is the act that conforms to a rule such that the consequences of *everyone's* complying with that rule are better than the consequences of everyone's complying with any alternative rule.
            1. This version of the theory solves the problems we noted for an individualistic version:
                a. The consequences of *everyone* complying with the rule "don't bother to vote" would be disastrous.
                b. If *everyone* in the developed world were donating an equal proportion of their income to the best charities, the optimal proportion would be nowhere near as high as 98% (the extreme demand arises rather from any given individual's attempt to 'take up the slack' left by the non-donaters).

2. Consider, however, the rule "keep strictly to the highway code, and optimise on the assumption that everyone else will do the same." This rule would have excellent consequences *if absolutely everyone complied with it*. But it is very unforgiving: In practice, if I assumed that every motorist will definitely maintain lane discipline on roundabouts, and that every cyclist will go straight on at junctions unless (s)he has signalled otherwise, I would get into an awful lot of accidents.
    a. The point is to agree on a code that will make things better. But we need to be realistic. If we choose our code on the assumption that everyone will comply strictly with it, we are "imagining out of existence" some serious problems that we'll soon discover, when we start implementing the rules.

iii. Society-wide majority-compliance rule-consequentialism: The right act is the act that conforms to a rule such that the consequences of *most people's complying with that rule most of the time* are better than the analogous consequences for any alternative rule.
    1. This theory seems to give the right results in our examples of deciding whether or not to vote, and deciding whether or not to abide strictly by the Highway Code and assume that all others will do likewise.

c. Objection to *any* 'compliance' version of rule-consequentialism: This theory collapses into act-consequentialism
    i. Consider any case in which act-consequentialism would recommend breaking the rule-consequentialist's rules: for instance, if the rule-consequentialist accepts the rule "don't lie", a particular case in which telling a lie would lead to better consequences than not telling the lie.
    ii. Given that such cases sometimes occur, acting in accordance with the alternative rule "don't lie except when the consequences of lying are better than the consequences of not lying" would lead to better consequences than obeying the existing rule "don't lie".
    iii. More generally: the consequences of obeying the rule "do what the act-consequentialist says" are – obviously – better than the consequences of obeying any alternative rule.
    iv. Reply: This is a problem *only* for a 'compliance' version of rule-consequentialism, not for an 'acceptance' version...

d. *Acceptance* rule-consequentialism: focuses on the consequences of *accepting* the rule.
    i. The point is that accepting a rule involves more than just *doing what the rule says*. It also involves:
    ii. Guiding one's decision-making by appeal to the rule.
        1. Accepting a very complicated rule, or one that is very complicated to apply in practice, would normally have bad consequences, because one would take a very long time to reach decisions, even on trivial matters.

2. Preview of Lecture 8: Notice that this is *not* the same as just 'doing what the rule says'. The distinction here is that between *a criterion of the right* and a *decision procedure* (on which more in Lecture 8).

iii. Making it publicly known that one generally makes decisions by appeal to the rule.
    1. Accepting a rule that permitted breaking promises whenever that would lead to a better outcome would probably have bad consequences overall, because others would lose trust in your promises (you would effectively lose the – very useful – ability to *make* promises).
    2. If doctors accepted a rule that permitted organ harvesting, this would (given the actual state of human nature, in which people care more for their own life than for the lives of strangers) have the bad consequences we noted above: both patients and visitors would tend to avoid doctors.
    3. These effects are "expectation effects".
    4. Actually, expectation effects are very likely to already be effects just of *complying with* the rule. (That's why the act-consequentialist *is*, in many cases, able to account for the wrongness of e.g. lying.) So *these* sorts of effects might not give rule-consequentialism any advantage over act-consequentialism.

iv. Training one's moral psychology to take the new rule into account.
    1. This takes time and effort. That time and effort amounts to a "transition cost" of accepting any given new rule (generally higher for rules that are more complicated, and/or less in keeping with one's existing moral psychology).
    2. One is likely to make mistakes during the process of transition, if the new rule is very different from the principles one is used to.
        a. For example, if we tried to train ourselves to accept a rule requiring strict impartiality between family and strangers, we would probably fall short of obeying the rule on numerous occasions – it's *incredibly* psychologically difficult to be completely impartial, given the kind of creatures we are. That in turn would have bad consequences – perhaps in terms of constant guilt feelings, and/or perhaps in terms of making us disenchanted with the whole enterprise of morality.

e. Defenders of rule-consequentialism argue that *this* version of their theory does not collapse into act-consequentialism, because, as objectors to act-consequentialism have long pointed out (and as we saw above), the consequences e.g. internalising an act-consequentialist criterion in one's moral psychology might very well be so bad as to justify the occasional utility sacrifice that is involved in sticking to the 'rules' (e.g., the sacrifice of four lives that is involved when a doctor declines to harvest organs from healthy individuals for transplant).

f. Acceptance of the rule *by whom*?

i. As for 'compliance' versions of rule-consequentialism, we can distinguish between versions that focus on the consequences of the *agent's* accepting the rule, and those that focus on the *universal or majority* acceptance of the rule in the society as a whole.
ii. As in the case of compliance theories, the "society-wide majority compliance" version of 'acceptance' rule-consequentialism appears to yield the most intuitively plausible verdicts on the cases that we might apply the theory to.

12. Objections to any form of rule-consequentialism
    a. Objection: Rule-consequentialism is incoherent
        i. The objection: rule-consequentialism is supposed to be motivated by an overarching commitment to maximising the good. But in that case, it is paradoxical for the advocate of rule-consequentialism to insist that any particular act that would maximise the good is nonetheless wrong (merely on the ground that *widespread acceptance of a rule permitting* this act would not maximise the good).
        ii. Reply: The rule-consequentialist's motivation need *not* be an overarching commitment to maximising the good. There is an alternative motivation: the desire to (i) capture most or all of the moral principles we ordinarily accept and (ii) to provide those principles with a unified foundation. Rule-consequentialists whose only motivations are these alternative ones are immune to this objection.
            1. Doubts about this move: This does seem to give up on one very important motivation, though. (Cf. above: "It's natural to think that the *justification* of the rules has something to do with making things better.")
    b. Objection: Rule-consequentialism recommends an implausible form of "rule-worship".
        i. Example 1: Disaster
            1. Suppose (as rule-utilitarians often seem to agree) that widespread acceptance of the rule "don't tell lies" would generally promote the best available consequences. Then a prohibition on lying would be part of the rule-consequentialist's moral code. But in some cases, telling the truth would lead to 'disaster'.
                a. Example 1: Kant's 'murderer at the door'
                b. Example 2: The malevolent intruder in the nuclear bunker
            2. Reply: Any plausible version of rule-consequentialism will also include a rule requiring one to prevent disaster whenever possible. So it won't have the implausible implication that one should tell the truth even to the murderer/intruder.
        ii. Example 2: The sheriff again
            1. In the notorious example of the sheriff and the innocent scapegoat, rule-consequentialism recommends sticking to the rule whose general acceptance by the majority of people would overall lead to the best consequences, despite the known fact that in this instance

breaking that rule would lead to better consequences. But rule-consequentialism provides no *rationale* for preferring compliance with the rule in such a case. To comply with it is simply to manifest an irrational rule-worshipping tendency.

2. Rule-consequentialists are likely to be unimpressed by this demand for a further rationale. It is, they will insist, just *true* that the sheriff ought not to hang the innocent man, and it is a virtue of the rule-consequentialist theory that it agrees with that verdict. Explanations have to stop somewhere.

c. Further reading:
   i. SEP, "Rule-consequentialism"
   ii. Hooker, "Ideal code, real world", esp. chapters 1, 3, 4

13. Mill's version of consequentialism
   a. Q: Is Mill an act-consequentialist, a rule-consequentialist, or what?
   b. Note that the issue under discussion, in this lecture, is not what makes one state of affairs *better* (either for a given person *or* overall) than another. (As we've seen, Mill's answer to that is the utilitarian *theory of the good*: that the relevant quantity is the sum of happiness/pleasure, net of unhappiness/pain, summed over all people.) Rather, our current question is *what makes acts right/wrong* (the 'criterion of right action'). So the question is what Mill says about the latter.
   c. Mill's initial statement of his view on rightness/wrongness is:
      i. "Acts are right in proportion as they tend to promote happiness, wrong in proportion as they tend to promote the reverse of happiness." (U 2.2)
   d. Comments on this first quote:
      i. Taken literally, this does not cleanly fit *any* of the types of consequentialism we canvassed above. It seems to employ a notion of rightness according to which rightness itself is a matter of degree: one action can be *more right* than another, by promoting happiness *more effectively*; most acts are right to some degree and also wrong to some degree.
      ii. However, many writers (e.g. Crisp) simply dismiss this aspect of Mill's use of 'right', and take the above quote as evidence that Mill's view is that of a maximising act-consequentialist.
   e. However, in chapter 5, Mill offers a much more complex account of what it takes for an action to be wrong:
      i. An act is *wrong* iff it "ought" to be *punished* in some way.
         1. "We do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, then by the opinion of his fellow-creatures; if not by opinion, then by the reproaches of his own conscience." (U 5.14)
         2. Besides actions that are wrong, "There are other things, on the contrary, which we wish that people should do, which we like or admire them for doing, perhaps dislike or despise them for not doing, but yet admit that they are not bound to do; it is not a case of moral obligation; we do not blame them, that is, we do not think that they are proper objects of punishment. ... [W]e call any conduct

wrong, or employ, instead, some other term of dislike or disparagement, according as we think that that person ought, or ought not, to be punished for it." The question is whether "we would wish to see the person whom it concerns, compelled, or only persuaded and exhorted, to act in that manner." (U 5.14, cont'd)

ii. This 'punishment' could take the form of:
    1. Legal punishment, e.g. imprisonment or a fine.
    2. The disapproval of others, as e.g. when one person treats another rudely.
    3. Unpleasant feelings due to having acted against one's own conscience, as e.g. when one commits some (moral or legal) 'crime' that goes undetected.

iii. But presumably Mill thinks that whether an action "ought" to be punished depends simply on whether it *would increase overall utility* to punish it.
    1. Mill recognises the prevalence of several conflicting theories of when punishment is 'appropriate'.
        a. "There are some who say, that it is unjust to punish any one for the sake of example to others; that punishment is just, only when intended for the good of the sufferer himself. Others maintain the extreme reverse... Mr Owen, again, affirms that it is unjust to punish at all; for the criminal did not make his own character... All these opinions are extremely plausible,; and so long as the question is argued as one of justice simply, without going down to *the principles which lie under justice and are the source of its authority*, I am unable to see how any of these reasoners can be refuted." (U 5.28; emphasis added)
    2. But it's clear from elsewhere in chapter 5 that these "principles which lie under justice" are principles of utility.
        a. E.g. On conflicting principles of what counts as fair remuneration for work involving unequal talents, Mill says: "Each [such principle], from his own point of view, is unanswerable; and any choice between them, on grounds of justice, must be perfectly arbitrary. *Social utility alone can decide the preference*." (U 5.30; emphasis added)
        b. On disputes between rival principles of fair taxation: "From these confusions there is no other mode of extrication than the utilitarian." (U 5.31)

f. Urmson advocates a *rule-utilitarian* interpretation of Mill: See Urmson, 'The interpretation of the moral philosophy of J S Mill', The Philosophical Quarterly', 3 (10), 1953

14. Deontological theories
    a. As we've seen, rule-consequentialists attempt to derive familiar moral principles from some criterion that (somehow!) includes an appeal to goodness of outcomes.
    b. An alternative approach to ethics takes such principles to be *fundamental*.

i. Advocates of this approach do not think that the principles in question need, or can be given, any deeper justification.
ii. Instead, they attempt to determine which are the right principles simply via considerations of the extent to which they match our intuitive moral judgments about particular cases.
iii. This approach is associated with (inter alia) the work of W. D. Ross, Judith Jarvis Thomson and Frances Kamm.

c. Ross's pluralism (W D Ross, *The right and the good*, esp. chapter 2)
i. There are multiple 'prima facie duties':
1. Fidelity
2. Reparation
3. Gratitude
4. Justice, in the sense of proportionment of happiness to virtue
5. Beneficience
6. Self-improvement
7. Non-maleficience
ii. The principles of prima facie duty are "self-evident", meaning that they do not require *proof* (although they may not be *immediately* obvious).
iii. In any given situation, what one is morally required to do (one's "duty proper") is determined by a process of weighing up and balancing among the various prima facie duties that apply to that situation. (Analogy: adding up component physical forces, to determine the resultant.)
iv. E.g. the decision as to whether to lie to the murderer involves a conflict of the prima facie duties of fidelity and beneficience.
v. There are no higher-order principles to resolve conflicts among prima facie duties – just use your "judgment".

d. A modern non-consequentialist project: the search for more precise principles
i. Ross's pluralism seems not to say as much as there is to be said on the issue of *when* one of his 'principles of prima facie duty' prevails over another.
ii. In an attempt to improve on this, modern non-consequentialists appeal to a number of thought-experiments in order to identify the correct, more complex, principles.

15. The evolution of more complex principles in response to counterexamples: A case study
a. 0^{th} principle (a utilitarian one, approximately): Minimise the number of lives lost.
b. Counterexample: Organ harvesting.
c. First principle: *Don't kill.*
d. Counterexample: Original Trolley Case
i. A runaway trolley is careening down a track. Trapped on the main track, ahead of the trolley, are five people, who will be run over and killed if nothing is done. A bystander has the ability to flick a switch that would divert the trolley onto a side-track. There is, however, one person stuck on the side-track, who will be killed if the switch is flicked.
ii. Should the bystander flick the switch?
e. Majority verdict on the Original Trolley Case, even among non-consequentialists: It is (at least) *permissible* to flick the switch.

i. This verdict conflicts with a "don't harm people" principle. (If the bystander does nothing, *he* does not kill the five. But if he flicks the switch, *he* does kill the one.)

f. Second principle: Don't kill, unless you can cause significantly more good by doing so.

g. Counterexample: 'Fat man on the bridge' case
   i. A runaway trolley is careening down a track. Trapped on the main track, ahead of the trolley, are five people, who will be run over and killed if nothing is done. A bystander is standing on a bridge, between the trolley and the five people; next to him is a very fat man. His only way of stopping the trolley is to push the fat man off the bridge, into the path of the trolley. If he does, the fat man will die.

h. Majority verdict on the Fat Man case (among non-consequentialists, anyway): It is *impermissible* to push the fat man.
   i. No version of our revised principle can capture both the majority verdict in the Original Trolley Case and that in the Fat Man case, since the numbers of people involved are the same in those two cases.

i. Third principle: The Doctrine of Double Effect (DDE)
   i. The DDE is based on a distinction between *intending* and *merely foreseeing* certain effects of one's actions.
   ii. A non-moral example, to illustrate the distinction: Suppose that I water my plants, in order to help them grow. I *intend* to help my plants grow. I may also *foresee* that another consequence of my action is that my cat drinks the water leaking out of the bottom of the plant-pots; but it doesn't follow that I *intend* that further consequence.
   iii. The DDE: One may not *intend* harm to others. One may, however, take actions that one merely *foresees* will cause harm.
   iv. Real-life applications of the DDE
      1. Application to medical ethics: It is generally accepted that a doctor may prescribe pain-killing medicine that she foresees will also shorten the patient's life, provided that the *intention* is pain relief, not euthanasia.
      2. Application to just war theory: Currently dominant theories of just conduct in warfare permit taking actions that one *foresees* will result in the death of innocent civilians (for example, bombing a munitions factory that is located close to a residential area), but prohibit actions that are *intended* to bring about the death of innocent civilians (for example, in order to terrorise the enemy population into surrender).
   v. How the DDE is supposed to deal with the first two trolley cases
      1. DDE on the original trolley problem: The bystander *foresees* that flicking the switch will result in the death of the person on the side-track, but he does not *intend* that consequence. (All he *intends* is that the trolley be diverted away from the main line.)

2. DDE on the Fat Man case: If he pushes the fat man off the bridge, the bystander must *intend* that the fat man die, since the fat man's collision with the train is a *necessary means* to saving the five – the latter being the whole point of the action. Therefore pushing is impermissible.

vi. An initial worry about the DDE
   1. The intend/merely foresee distinction is not clear enough to bear the weight that the DDE theorist seeks to put on it. Why not say, for instance, that the bystander *intends* only that the fat man stop the train, and that he *merely foresees* that this will cause the fat man's death?
   2. But even if this worry can be satisfactorily answered, other cases seem to show that the DDE cannot be consistent with commonly held intuitions in all cases.

vii. Counterexample to the DDE: Loop case
   1. This case is like the Original Trolley Case, except that the 'side-track' loops round and rejoins the main track, so that the diverted trolley would still run over the five on the main track – but for the fact that the body of the one person who is stuck on the loop would stop the trolley before it completes the loop.
   2. Majority verdict: There is no morally relevant difference between the Loop Case and the Original Trolley case. It is permissible to redirect the trolley onto the loop.
   3. But on any account of the intend/foresee distinction according to which one intends to kill the fat man when one pushes him off the bridge, one surely also intends to kill the one person when one diverts the trolley onto the Loop. Therefore the DDE cannot agree with the majority verdicts in both cases.

viii. Kamm's reply: The Doctrine of Triple Effect (DTE)
   1. As the Doctrine of Double Effect is based on a distinction between intending and merely foreseeing, the Doctrine of Triple Effect regards a further distinction as also being key: that between taking an action *because of* some foreseen effect, and taking it *in order to bring about* that effect.
   2. Illustrative example: The Party
      a. We *intend* to throw a party in order to have fun. We *foresee* though that this will result in a big mess, and we will not have a party if we will be left to clean up that mess by ourselves. However, we also foresee that if we throw the party, our friends will feel indebted to us and this will cause them to help clean up. Hence, we throw the party *because* we believe that our friends will feel indebted and because they will help us clean up. But we do not give the party *in order to* make our friends feel indebted or to cause them to clean up for us.

3. The Doctrine of Triple Effect: One may not act *in order* to bring about harm. But one may act *because of* a foreseen harm that one's action would cause.
    a. "Triple effect" because we now have three categories: harms that are *intended*, harms that are *merely foreseen*, and harms that are foreseen and in addition that one acts *because of* (but does not intend).
    b. Kamm argues that in the Fat Man case, the bystander would be pushing the Fat Man *in order* to cause the trolley to run him over, but that in the Loop case, the bystander would be flicking the switch merely *because* the trolley will then hit the one person on the side-track.
        i. This difference is supposed to be present because in the Loop case, merely diverting the trolley onto the loop track averts the *original* threat to the five. It's true that it simultaneously creates a new and potentially equally potent threat (of the trolley looping round and hitting the five after completing its loop), but that is a different threat. And there would be no point in averting the original threat if the new and equally bad threat would indeed materialise. What the presence of the One does is to defeat this new threat. The switch-flicker acts *because* the trolley's hitting the One will defeat the new threat. But it does not follow, according to Kamm, that the switch-flicker *intends* for the trolley to hit the One, any more than the party-thrower *intends* to get his friends to feel guilty.
        ii. In contrast, in the Fat Man case, no such appeal to defeating a new threat can be given: there is no switch, in this case, from an old to a new threat. The bystander would, in that case, be pushing the fat man *in order to* get the trolley to hit him.
4. Kamm's book "Intricate ethics" discusses both the formulation and application of this "Doctrine of Triple Effect", and some potential problems for the doctrine, in much more detail.

16. Summary: In this tradition of ethical theorising,
    a. The key tool is the thought-experiment.
    b. The "data" are one's considered intuitive judgments on the particular cases described in the thought-experiments (e.g., on the various Trolley cases).
    c. One seeks principles that successfully reproduce all of these judgments (or as many as possible).
    d. The *intrinsic* plausibility of the principles themselves (i.e., aside from the extent to which they match intuitive particular-case judgments) is less of a concern.
17. Mill's criticisms of the "intuitionists" (Cf. U, chapter 1)

a. The non-consequentialist theorists we have discussed above (Ross, Thomson, Kamm) are writing *after* Mill. (Mill's Utilitarianism: 1861. Ross's The right and the good: 1930. Thomson and Kamm are present-day philosophers.)
b. The "intuitionists" Mill had in mind included Thomas Reid, William Hamilton, and William Whewell (master of Trinity College, Cambridge).
c. Still, Mill's key objections to "intuitionism" would apply equally to the above approach:
    i. The appeals to "moral sense" and "intuition" needed to make epistemological sense of the intuitionist's methodology – relying on sources of knowledge that are based neither on sense experience nor on reasoning (whether deductive or inductive) – are unacceptable.
    ii. The intuitionist approach is too conservative. Since it treats existing moral judgments as sacrosanct, it can never lead to moral progress.
        1. This is anathema to Mill's ambitions of social reform.
d. We will return later to the issue of whether these criticisms are fair – in particular, the extent to which Mill's own approach fares any better with respect to them (cf. e.g. the later discussion of reflective equilibrium).

Lecture 5: Moral motivation, moral epistemology and meta-ethical realism (I)

18. Moral motivation
    a. Mill's Chapter 2 has sketched his utilitarian theory of morality. Chapter 3 then takes up the task: what might *motivate* someone to act as this theory says?
    b. If nothing could provide the backing motivation, then it's pointless to put forward the theory.
    c. Biographical note: This concern seems to have had something to do with Mill's nervous breakdown, at the age of 20.
    d. The discussion of moral motivation more generally will also introduce us to the topic of *meta-ethics* (as opposed to *normative ethics*), involving questions of reality and truth in ethics and how 'moral truths', if they exist at all, might be known.
19. Mill on moral motivation
    a. Mill anticipates the claim that nothing could *motivate* someone to do as utilitarianism requires being raised as an objection to utilitarianism.
    b. Mill's reply is: Insofar as there is any difficulty of this sort now, that is only because utilitarianism is not now the 'customary morality'.
        i. But this isn't to say it couldn't become so.
        ii. And once it does, the relationship to motivation will be the same as for any other moral principle.
    c. Mill's account of moral motivation in general
        i. There are both external and internal 'sanctions' that, in the vast majority of individuals, cause the individual to *want* to behave as morality dictates.
            1. External sanctions: punishment by God and/or disapproval from fellow humans. "The hope of favour and the fear of displeasure, from our fellow-creatures or from the Ruler of the Universe."
            2. Internal sanctions: "a feeling in our own mind, a pain, more or less intense, attendant on violation of duty, which in properly cultivated

natures rises, in the more serious cases, into shrinking from it as an impossibility."

ii. It is theoretically possible that someone might just not care about the disapproval of others, and might just not experience the "feeling in his own mind" that Mill writes about (i.e., that he might not have a conscience).
    1. Mill simply concedes that such a person will have no motivation to conform to morality. (The internal sanctions of morality "have no binding efficacy on those who do not possess the feelings it appeals to."
    2. Again, this applies to any other morality just as much as it does to utilitarianism.

iii. Mill thinks that, if anything (and contingencies of which morality is currently 'customary' aside), the issue of motivation is *easier* for utilitarianism than for alternative moralities.
    1. This is because utilitarianism "harmonises with" some of our natural feelings, in a way that an arbitrary morality would not.

d. Alternatives to Mill's account of moral motivation
    i. Note that Mill's account is focussed on explaining how the motive of *wanting pleasure and the absence of pain for oneself* might lead to motivation to do what one believes to be the morally right thing. (Thus his discussions of internal and external sanctions focus on the *displeasurable consequences of* violating what one takes to be one's duty.)
    ii. This is rationalised by the hypothesis that Mill subscribes to *desire egoism*: Each person has (final) desires only for his or her own happiness.
        1. Crisp (in his "Mill on virtue as part of happiness") also attributes desire egoism to Mill (although not for this reason).
        2. Is desire egoism true? Some doubts:
            a. Can't I have final desires at least for the happiness *of other people?*
                i. Mother Theresa-style examples
                    1. The desire egoist's reply: Mother Theresa is doing it for the 'warm glow' (a positive internal sanction). She doesn't have an *ultimate* desire for the welfare of others.
                    2. This is not very plausible, but it's hard to disprove directly, *in these cases*.
                ii. The soldier who throws himself on the grenade
                    1. The point of this example is: Clearly the soldier isn't going to get any 'warm glow', or at any rate, not enough (at most, he'll get a split second's worth).
                    2. A *possible* reply: The soldier would feel so guilty, if he didn't do this, that his continued life would be worse than dying.

3. Again, it is perhaps not very plausible that this is the only/main motive, but hard to disprove.

iii. The mother who chooses that her children's lives go well while she is tricked into believing she made the opposite choice, rather than vice versa

1. This example seems conclusively to refute desire egoism.

b. Can't I have desires whose object is not directly *anyone's* happiness, e.g. desires *to do the right thing*?

i. The reason I desire not to steal from old ladies' homes is because I think that would be seriously morally wrong, and I desire not to commit serious moral wrongs. (I would *still* have that desire even if, somehow, you could arrange things so that I believed I would not suffer guilt.)

iii. Alternative accounts of moral motivation, eschewing desire egoism (See also SEP, "Moral motivation")

1. We can be brought to have desires directly for others' welfare. (As in the case of the mother.)
   a. The further this process goes, the closer we get to utilitarian motivation. (Utilitarianism is the logical extreme of this phenomenon.)
   b. This "direct desires for others' welfare" approach is really just a small deviation from Mill's view that we can be brought into a state such that the 'internal sanctions' motivate us to act so as to enhance others' welfare (but avoids the counterexample of the mother).
2. Most of us have a standing desire to do the right thing (or at least: not to do things that are *seriously* wrong). That is why, when we come to believe that some act would be wrong, we form a derivative desire not to perform that act.
   a. Smith's objection: "moral fetishism"
      i. Example: You are considering whether to club an old lady over the head with a baseball bat in order to steal her purse. You decide not to, because you know it is wrong and you don't want to do wrong.
      ii. Smith's reaction: being motivated by such an abstract "desire not to do wrong", rather than e.g. by concern for the old lady, is a moral *vice*. The *virtuous* person cares *directly* about the individual acts that she thinks are wrong, not only "indirectly" (i.e. via her belief that they would be moral wrongs).
3. Most of us are just wired up in such a way that, if we come to believe that some act is wrong, we immediately form a corresponding desire not to perform that act.

    a. How this applies to the example of the old lady: We know that clubbing old ladies over the head with baseball bats is wrong. Our psychology is such that a causal result of having that belief is that we also have the *desire not to club old ladies.*
    b. This is subtly different from the "desire to do the right thing" setup. The present version is immune to Smith's criticism.
    c. A worry: does this (still, i.e. in common with the previous suggestion) make moral motivation too contingent, too accidental? What if someone didn't happen to *have* the relevant desires or desire-forming tendencies?
4. Interlude: the Humean theory of motivation
    a. Humean theory of motivation: motivation always requires, in addition to belief, the presence of a desire. Belief cannot motivate *on its own*.
    b. Example: Your believing that the lecture starts at 12pm will not generate any motivation to get to Exam Schools for 12 unless you also have a *desire* to go to the lecture.
5. Anti-Humean theory of motivation: Some beliefs, notably beliefs that some act is wrong, are enough *on their own* to generate a motivation not to perform that act (that is, no accompanying 'desire' *at all* is needed – not a *desire* not to club old ladies, and not a *desire* to avoid doing moral wrong either).
    a. This approach is associated with (although it *need* not accompany or be accompanied by) the idea that morality can be derived from 'reason'/'rationality' alone.
    b. It perhaps captures some elements of moral phenomenology better than any 'desire'-based account.
        i. Example: You have promised to meet your auntie for dinner, but you don't really enjoy her company. A friend invites you on a pub crawl. You decide to meet your auntie, because you have promised. But it might not feel quite right to say that you *want* to keep your promise, or to meet your auntie, or to do the right thing. It might feel more accurate to say instead: What I *want* to do is go on the pub crawl, but I know that I *ought* to keep my promise, so I'll do that.
    c. Motivational internalism
        i. The claim that a moral belief on its own (i.e. independent of any accompanying desire) can generate moral motivation is known as *moral internalism*.
        ii. Internalists support this claim by noting that there would at the very least be *something very odd* about someone who believes that (say) rape is wrong, but who has *no* corresponding motivation not to rape.

1. Note that this requires much more than merely that his moral motivation be overridden by some other motivation, or be defeated by e.g. weakness of will: the idea is that the character in question has *not even one component* of motivation in favour of not-raping.

iii. The internalist claim is that this is not merely odd, but actually conceptually impossible: according to internalism, part of what it is to *have* the belief that rape is wrong is to have at least some element of corresponding motivation.

iv. Externalists, on the other hand, say that such cases of motivation failure are not conceptually impossible, just (thankfully) extremely rare.

1. But (according to externalists) they do exist!
2. Arguable examples: psychopaths. Psychologists report that e.g. serial killers often say that they perfectly well *understand* that it is morally wrong to kill people, but that they *just don't care* about that.
3. Internalists have to argue that such psychopaths actually don't believe that it's wrong to kill, despite what they (presumably sincerely) say.

iv. As far as defending utilitarianism from the charge that it cannot motivate is concerned, Mill could equally have subscribed to *any* of these theories of moral motivation.

1. 'Desire to do the right thing' version: by engaging in moral reasoning of the sort in Mill's Utilitarianism and/or of the sort we have been engaging in in the past few lectures, we might be brought to form the moral *belief* that utilitarianism is true, and hence e.g. that we *morally ought* to give lots of money to the most cost-effective charities. Combined with a standing desire *to do what we morally ought to do*, this generates *motivation* to give to charity.
2. 'Causal link between moral beliefs about an action and desires about that action' version: We come to the belief that we ought to give lots of money to charity in the same way as above. A causal consequence of forming that belief is that we also form the *desire* to give lots to charity.
3. Anti-Humean version: We come to the belief that we ought to give lots to charity, again in the same way. That is enough *on its own* to motivate us to give lots to charity.
    a. Mysteriously?
    b. Or: Because motivation was *already* involved, throughout the process of moral belief-formation (e.g. in our deliberations about the 'truth' of utilitarianism).

c. This is less mysterious if we don't think of ethical 'truths' in the sense of truths independent of us (on the model of e.g. physical truths about the world), but rather as themselves being *codifications* of certain aspects of our attitudes, e.g. attitudes of approval/disapproval towards certain actions).
    i. This leads on to the discussion of realism/non-cognitivism/etc., next week.

20. Moral epistemology: Mill's "proof" of the "principle of utility"
    a. In Chapter 1, Mill warns us against demanding a *proof* of the fundamental principle of utilitarianism, in any strict sense of 'proof'. This is because it's supposed to be a *fundamental* principle. And fundamental principles are the things that *other* propositions are proved *from* – precisely because they are fundamental, they cannot themselves be proved.
        i. But, according to Mill, this *doesn't* mean that the choice of first principle is arbitrary, or that the correct 'first principle' cannot be known. "It is evident that [such proof as the utilitarian theory is susceptible of] cannot be proof in the ordinary and popular meaning of the term. Questions of ultimate ends are not amenable to direct proof. ... We are not, however, to infer that its acceptance or rejection may depend on a blind impulse, or arbitrary choice. There is a larger meaning of the word proof... The subject is within the cognisance of the rational faculty, and neither does that faculty deal with it solely by way of intuition. Considerations may be presented capable of determining the intellect either to give or withhold its assent to the doctrine; and this is equivalent to proof." (U 1.5)
    b. Chapter 4, titled "Of what sort of proof the principle of utility is susceptible", takes up the task of providing these "considerations" to "determine the intellect" to accept Mill's principle of utility.
    c. Reminder: remember the distinction between something *as a means* (= having an "instrumental" desire for it) and desiring it *as an end* (having a "final" desire for it)
        i. The example used earlier: Catching the 6 o'clock train to Cardiff

21. An initial exposition of Mill's "proof"
    a. Structure of the "proof"
        i. Step 1: Happiness is desirable.
        ii. Step 2: Nothing other than happiness is desirable.
        iii. Step 3: "The happiness of all" is a good to "the aggregate of all persons".
    b. Exposition of Step 1
        i. Mill seeks to establish that happiness is desirable via the *empirical* observation that it is in fact desired:
            1. "The only proof capable of being given that an object is visible, is that people actually see it. The only proof that a sound is audible, is that people hear it: and so of the other sources of our experience. In like manner, I apprehend, the sole evidence it is possible to produce that anything is desirable, is that people do actually desire it. If the end which the utilitarian doctrine proposes to itself were not, in theory and in practice, acknowledged to be an end, nothing could

ever convince any person that it was so. No reason can be given why the general happiness is desirable, except that each person, so far as he believes it to be attainable, desires his own happiness. This, however, being a fact, we have not only all the proof which the case admits of, but all which it is possible to require, that happiness is a good: that each person's happiness is a good to that person..." (U 4.3)

c. Exposition of Step 2
    i. Step 1 depends on accepting the methodological principle that if people do in fact desire something (and perhaps if this desire survives critical scrutiny), then that thing is desirable.
    ii. But as Mill is all too aware, it certainly *seems* to be the case that people actually desire other things too, besides happiness.
        1. Examples: money, virtue.
    iii. Why aren't they desirable too, then? And if they are, isn't this a counterexample to Mill's claim that only happiness is desirable 'as an end'?
    iv. Mill's answer:
        1. Anything that someone desires as an end is a "part" or "ingredient" of (*not* a "cause of" or "means to") that person's happiness.
            a. What does Mill mean by a "part" of happiness? This is pretty obscure! More on this below.
        2. Therefore nothing is desired as an end "apart from" happiness, i.e. nothing that is not either happiness itself or some part of happiness.

d. Exposition of Step 3
    i. Mill says extremely little here. The passage quoted in connection with Step 1 above continues "...and the general happiness, therefore, a good to the aggregate of all persons." That is all that *Utilitarianism* has to say on the matter. (As we noted in Week 3, it's not enough – it gives no reason to prefer utilitarianism to e.g. prioritarianism/egalitarianism.)

22. On what this "proof" is supposed to be a "proof" of
    a. Note that it's a proof of the "principle of utility" in the sense of *theory of the good*; nothing at all is said here about criterion of right action.
    b. Mill intends it to be a proof of his theory of the overall good, although, as we've seen, he only really argues for the "theory of well-being" part of this (i.e., for hedonism).
23. Criticisms of Mill's "proof", and replies
    a. Criticism of Step 1
        i. 'Desirable' and 'visible' are *not* analogous. Their meanings differ in a crucial way that undermines Mill's argument.
            1. "[With the visible/desirable analogy, Mill] pretends to prove that good means desired.

               "Well, the fallacy in this step is so obvious, that it is quite wonderful how Mill failed to see it. The fact is that desirable does not mean able to be desired as visible means able to be seen. The desirable means simply what *ought* to be desired or *deserves* to be desired;

just as the detestable means not what can be but what ought to be detested and the damnable what deserves to be damned. Mill has, then, smuggled in, under cover of the word desirable, the very notion about which he ought to be quite clear. Desirable does indeed mean what it is good to desire; but when this is understood, it is no longer plausible to say that our only test of *that*, is what is actually desired." (Moore, *Principa Ethica*, ch. 3)

ii. Reply on behalf of Mill:
    1. Mill was not making the (as the critics have been quick to point out, obviously false) claim that desirable *means* 'able to be desired', as visible means 'able to be seen'.
    2. On reflection, it's clear that he *can't* have meant that, since if he had then he should have claimed that his argument was a strict proof, and we have seen him explicitly deny that 'proof' in any strict sense is impossible here.
    3. His point is rather that since we all find that we do desire happiness and that we continue to desire it on reflection, we already accept that happiness is desirable, and we need no further proof of that claim.
    4. The analogy between 'desirable' and 'visible' is in large part a showman's flourish, inserted for rhetorical effect but not to be taken too seriously.

b. Further discussion and criticism of Step 2
    i. Mill on virtue and money
        1. Utilitarianism itself actually entails that virtue "is to be desired disinterestedly, for itself". This is because a mind is "not in the state most conducive to the general happiness, unless it does love virtue in this manner – as a thing desirable in itself, even though, in the individual instance, it should not produce those other desirable consequences which it tends to produce, and on account of which it is held to be virtue". (U 4.5)
        2. Once virtue *has* come to be desired "for its own sake", it has become part of the person's happiness: "What was once desired as an instrument for the attainment of happiness, has come to be desired for its own sake. In being desired for its own sake it is, however, desired as *part* of happiness. The person is made, or thinks he would be made, happy by its mere possession". (U 4.6)
        3. The analogy to money: "There is nothing originally more desirable about money than about any heap of glittering pebbles. Its worth is solely that of the things which it will buy; the desires for things other than itself, which it is a means of gratifying." But over time, people come to want money for its own sake: "It may, then, be said truly, that money is desired not for the sake of an end, but as part of the end. From being a means to happiness, it has come to be a principle ingredient of the individual's conception of happiness." (U 4.6)

ii. Objection: It's question-begging to insist, as Mill *apparently(?)* does, that just because virtue/money/etc is desired for its own sake, it is desired as "part of" happiness, rather than as an independent good.
iii. Possible reply, on behalf of Mill: Mill doesn't assume that this is something the objector will already agree with (so he's not begging the question, as such); he's rather making a *further psychological claim*, namely that what is in fact going on in cases of desire for money/virtue "for their own sakes" is that the person "is made, or thinks he would be made, happy by their mere possession", and that is what explains the presence of the desire.
iv. Counter-objection:
    1. If this is indeed Mill's hypothesis, it would be more accurately stated as: The person thinks that possessing virtue/money would *cause* happiness in him, thus these are actually cases of desiring virtue/money as a means (to happiness), not as an end, after all.
    2. But *that* hypothesis can be put to the test in a thought-experiment: would the desire for money/virtue still remain, even if the belief that it would cause happiness were removed?
        a. E.g. a (somewhat fanciful) example: Suppose that the man who "desires virtue for itself" can choose either to be virtuous *but be brainwashed into thinking that he's not*, or vice versa.
        b. If such a man chooses to *be* virtuous, *or even thinks that there are two sides to be weighed up here*, then his desire for virtue appears *not* to be 'instrumental', even in the sense of being dependent on a belief that being virtuous will make him happy.
            i. In genuine cases of instrumental desire, removal of the belief that the object of that desire will increase the chance of getting the object of final-desire *causes the instrumental desire itself to evaporate altogether.* (The 6pm train to Cardiff, again.)
v. Conclusion:
    1. Mill's attempt to claim that virtue, when desired "for itself", is desired "as part of happiness" appears indefensible.
    2. He would have done better to say that a final desire for virtue is a misguided desire, since happiness is the only thing that is *really* desirable as an end in itself.
    3. But Mill is prevented from taking this course by his empiricist scruples: he thinks, as he states in his argument, that what is actually desired is "the sole evidence" one could have regarding what is desirable.

24. Moral epistemology and meta-ethics more generally
    a. The difficulties that Mill encounters in giving any 'proof' of a theory of morality that might persuade anyone who is not already a utilitarian are symptoms of much deeper problems surrounding the epistemology of moral theory (i.e., the theory of

how one can know which morality is correct) – issues that Mill does not explicitly face up to.

b. Mill seems to assume a straightforward *realism* about morality – that there are moral truths, out there in the world and independent of what anyone's moral opinions are, just as for physical facts/truths.
c. Many people find this sort of realism implausible on metaphysical grounds.
d. Non-realist accounts
    i. Put off by the 'spookiness' of moral realism, many people have been drawn to alternative accounts, according to which moral talk is not to be taken literally.
    ii. E.g. the "simple subjectivist" account: "X is wrong" just means "I disapprove of X".
    iii. Such non-realist accounts might seem to lead to fewer problems of both moral motivation and moral epistemology, as well as being less metaphysically weird/spooky…
    iv. More on this in the next lecture...

Lecture 6: Interlude: meta-ethics

1. We have seen that in Mill's Chapters 3 and 4, he delves into the issues of
    a. Moral motivation: how might someone be *motivated* to act in accordance with the dictates of a moral theory?
    b. Moral epistemology: how can we *know* (or even: *have non-arbitrary beliefs regarding*) which moral theory is correct?
2. Both these questions turn on deep and difficult issues in "meta-ethics". Mill does not discuss these background issues at all.
    a. *Normative ethics* asks questions about what one *ought* to do, what determines whether an act is *wrong*, what makes a person's life *go well*, etc. (The italicised terms are *normative* or *evaluative* terms.)
    b. *Meta-ethics* asks questions about the status of the claims investigated in normative ethics. For example: When we say e.g. "murder is wrong", are we expressing an objective truth, or rather (e.g.) just giving voice to our emotional reaction to murder?
3. Aim of today's lecture: survey some of the possibilities, to give you the tools to think through for yourself whether you agree with Mill's treatment of the issues of motivation and epistemology.
4. The 'robust realist' account of morality
    a. There are moral properties (of goodness and badness, rightness and wrongness) that attach to acts, states of affairs etc: for instance, an action of *setting fire to a cat for fun* possesses the *property of wrongness*, just as my chair possesses the property of blueness.
        i. This is just to take moral talk literally:
            1. "My chair is blue": Standard approaches (as in your first-year logic course) interpret this sentence as asserting that a certain *property* (that of blueness) is possessed by a certain *object* (my chair).

2. "Setting fire to cats for fun is wrong": according to robust moral realism, the situation is just the same here.
    a. Non-realists have to say that even though the sentence structurally looks very much like an ordinary descriptive sentence, it actually works very differently.

b. These properties are just as real (and just as mind-independent) as physical properties such as shape, size, colour, scent etc.
c. However, whereas shapes/sizes/colours/scent/etc. can be seen/smelt/etc. – i.e., detected by means of ordinary senses whose physiology is fairly well-understood – moral properties cannot be detected by any of *those* senses.
d. This raises a crucial question: *how could we possibly know* which acts are right, and which wrong, on this account? ('Moral epistemology.')
e. The (realist's) usual answer: *moral intuition*.
    i. *Intuition* tells us either
        1. that causing pain, killing, telling lies, stealing, etc., are (at least *normally*) wrong, or
        2. that certain particular acts – that act of setting fire to cats, this act of lying to one's mother, etc. – are wrong.
    ii. This 'moral intuition' operates somewhat like a sixth sense. The usual five senses detect *natural* properties (redness, sweetness, squareness etc.); moral intuition detects *moral* properties.

5. Two key sources of dissatisfaction with the robust realist account (cf. Mackie, *Ethics: Inventing right and wrong*, ch.1, esp. sec. 9, 'The argument from queerness')
    a. Its metaphysics is weird, and extravagant: it postulates special moral properties that are no part of the furniture of the world according to natural science, and furthermore that don't seem to be required in order to explain anything we observe (even the fact that we make moral judgments!).
    b. Its account of moral epistemology is implausible: there is no account of how the "faculty of moral intuition" is supposed to work.
        i. Even if the realist's special "moral properties" did exist, note that even the realist doesn't think we are in *causal contact* with those properties.
6. Mackie's response: Error theory
    a. Mackie thinks that the realist is right about what our normal moral discourse *commits* us to: for a moral claim such as "abortion is wrong" to be true, it is necessary that there be a moral property of wrongness out there in the world, and that acts of abortion possess that property (just as, in order for "my chair is blue" to be true, it is necessary that there be a property of blueness, and that my chair possess that property).
    b. But Mackie thinks that in fact, no such moral properties exist.
    c. He therefore embraces an *error theory*, according to which our moral discourse is systematically mistaken.
    d. If we agree with Mackie on this: Where next?
        i. Mackie's error-theory raises the question of what we ought to do next. Normally, if one thinks that some body of claims is false, one stops making those claims. But surely it isn't a good idea just to *stop moralising*?

7. Other alternatives to the robust-realist account
    a. The simple subjectivist account: “Abortion is wrong” *just means* “I disapprove of abortion”.
        i. Objection to simple subjectivism: If simple subjectivism were correct, then when you say “Abortion is wrong” and I say “abortion is not wrong”, *we are not disagreeing* – I have not contradicted you. But in fact we are disagreeing. Therefore simple subjectivism is not correct.
    b. Culture-relative subjectivism: “Abortion is wrong” just means “abortion is frowned upon in my culture”.
        i. Objection to culture-relative subjectivism: This view cannot make sense of *cross-cultural* moral criticism.
            1. If a Nazi officer says “it’s OK to kill Jews”, what he says is *wrong*. But according to culture-relative subjectivism, he’s just saying “killing Jews is not frowned upon in my culture” – which may have been correct.
    c. A more sophisticated non-realist account: *Emotivism*
        i. Emotivists note that not all uses of language consist in making statements that are either true or false. One has
            1. Declarative sentences, e.g. “this chair is blue”. *These* are either true or false. But one also has, e.g.,
            2. Questions: “Did you go out last night?”
            3. Commands: “Shut the door.”
            4. Expressions of feeling: “Boo!”
        ii. None of the expressions in the last 3 categories is true/false.
        iii. The emotivists’ idea is that moral judgments are *expressions* of emotion/approval/disapproval.
            1. Saying “Abortion is wrong” is like saying “Boo to abortion!”
        iv. This avoids the objections to simple subjectivism and cultural relativism.
            1. *Expressing* one’s attitude is not the same thing as *stating* what one’s attitude is.
                a. The sentence “I disapprove of abortion” is truth-apt. The sentence “Down with abortion!” is not.
            2. On disagreement:
                a. It cannot happen, on the emotivist account, that two parties to an apparent moral disagreement must recognise that actually what the other person is saying is *true*.
                b. So there is no *positive* reason to think that emotivists must have a problem with moral disagreement.
                c. Further, it seems they have a fairly plausible account of what is going on in disagreement cases. If I say “boo to abortion!” and you say “hooray to abortion!”, there is a clear sense in which *our attitudes clash*.
8. Realist and non-realist accounts of moral motivation

a. Note that according to the realist account, moral judgments (e.g. "setting fire to cats is wrong") are simply *beliefs about the way the world is*, not e.g. expressions of desires.
b. This makes the phenomenon of moral motivation *prima facie* puzzling: ordinarily, at any rate, beliefs *on their own* do not generate any motivation (cf. the Humean theory of motivation, again).
    i. E.g. Your *belief that the lecture starts at 12 o'clock* will not motivate you to head to Exam Schools at 11.50 unless you also *want* to attend the lecture.
c. As we noted last week, some realists are happy to 'bite this bullet', and agree that your belief that setting fire to cats is wrong simply will not motivate you unless you happen also to have a relevant desire (e.g. the *desire not to do wrong*, or a desire not to set fire to cats that is caused, psychologically, by the presence of the moral belief) – noting, however, that in practice the vast majority of us do have this desire/desire-forming tendency.
d. Others think there is more to moral motivation than this.
e. This is another source of motivation for non-cognitivism.
    i. Note that there is no analogous 'puzzle of motivation' for non-cognitivists – there's no difficulty in understanding how the fact that one *disapproves* of (say) stealing leads to motivation not to steal.

9. Realist and non-realist accounts of moral epistemology
    a. Recall that the realist apparently had to postulate a (mysterious?) "faculty of moral intuition", in order to explain how we might get our moral beliefs to line up with the mind-independent facts regarding which actions were right/wrong.
    b. The non-cognitivist does not need any such mysterious appeal to intuition. There's nothing especially puzzling about how one can know, e.g., what one approves and disapproves of.
    c. This is a third source of motivation for non-cognitivism.
10. Trouble in the non-cognitivist camp
    a. We've seen several powerful motivations for non-cognitivism: one relating to metaphysics, one relating to motivation, and one relating to moral epistemology.
    b. But non-cognitivism in turn faces its own difficulties…
    c. A preliminary difficulty: Incompleteness
        i. Emotivism, for instance, says that the sentence "lying is wrong" functions to express disapproval of lying.
        ii. But such sentences can occur not only on their own, but also in embedded contexts – e.g. as part of "if lying is wrong, then getting your little brother to lie is wrong."
        iii. And it's clear that the subsentence "lying is wrong" does *not* express disapproval of lying in *this* context – the conditional sentence might e.g. be uttered by someone who has *no particular view* on whether or not lying is wrong.
        iv. So the emotivist theory is, as it stands, incomplete.
    d. A further worry: Incomplet*ability*

i. The emotivist might, of course, reply by adding to his theory, to provide an emotivist account of the meaning of "lying is wrong" when it occurs inside a conditional.
ii. But, for the reason already given, that will have to be a *different* meaning from the meaning that "lying is wrong" has when *un*embedded.
iii. And this is going to create a problem…

e. … Examples of moral reasoning: "Lying is wrong. And if lying is wrong, getting your little brother to tell lies is wrong. Therefore, getting your little brother to tell lies is wrong."
i. This appears to be a *valid argument*, in just the same way that a modus ponens argument about a purely descriptive matter would be valid.
ii. But if "lying is wrong" has a *different meaning* on the two occasions it occurs in this argument, then the argument ought to be guilty of a fallacy of equivocation.
iii. Example of the fallacy of equivocation:
1. My brother works at a bank.
2. If my brother works at a bank, then he works next to a river.
3. My brother works next to a river.
a. If 'bank' means 'financial bank' in (1) but 'river bank' in (2), then this is not a valid argument, despite the fact that *purely formally* it looks like an instance of modus ponens.

f. The basic problem here is that moral discourse seems to behave *structurally* just like descriptive reasoning. It turns out to be extremely difficult to explain this, on a non-cognitivist account of morality, for the sorts of reasons we have started to survey. (This is "The Frege-Geach problem" – of which more in the second year course! For the curious, try Schroeder' survey article 'What is the Frege-Geach problem?', *Philosophy Compass* 3/4, 2008.)
g. Aside: This is 'moral philosophy meets the philosophy of language'

11. Moral epistemology revisited (See SEP, "Reflective equilibrium")
a. Mill, as we've occasionally seen, wanted to avoid appeals to intuition in defending his theory: his project was to justify the theory as far as possible on an *empirical* basis.
b. And indeed, utilitarians often object against e.g. Thomson-Kamm-style deontological theorising that those theories place far too much weight on 'intuition', and as a result end up with excessively conservative systems of ethics.
c. On the other hand, it should be obvious that one cannot give any justification of any normative principle on an *entirely* empirical basis.
i. A complete description of how the world *is* does not on its own, logically, entail anything about how it *ought* to be.
d. In particular, we have seen (arguably?) that Mill's own attempt to defend even hedonism on a purely empirical basis fails.
i. Recall: Firstly, it presupposed desire egoism, which appears to be false. Secondly and more fundamentally, even if desire egoism were true, we might be left wondering whether the things that we do in fact find ourselves judging, even on reflection, to be valuable are *really* valuable, and, as Mill

would probably admit, an empirical approach has no resources with which to answer that question.

e. A different account of the methodological bone of contention:
    i. The real methodological issue between utilitarians and "intuitionists" concerns the relative weight they place on intuitions concerning fundamental explanatory principles (e.g. the principles of hedonism and/or of consequentialism) on the one hand, and on intuitions about which actions are right/wrong in particular cases on the other.
    ii. Several of the objections to act-consequentialism have the form: "This theory predicts that several clearly impermissible acts, in particular cases (e.g. that of the Sheriff), would be right". "Clearly" here means: "intuitively".
    iii. This illustrates the fact that the *general principles* we tend to find most intuitive are not consistent with the *particular-case judgments* that we tend to find most intuitive.
    iv. If we seek consistency, one or the other (or both) has to be modified.
    v. "Reflective equilibrium": the process of seeking consistency by revising one's initial judgments, about cases and/or general principles, in the light of tensions between them.
    vi. A majority of those working in normative ethics today take reflective equilibrium to be The methodology of ethics.
        1. So, in particular: everyone is appealing to "intuition", in some sense.
    vii. But this leaves open the disagreement over which intuitions (particular-case or general-principle) we should be most reluctant to revise.
    viii. It also leaves hanging the question of what, if anything, *justifies* us in treating intuitions as evidence in ethics.

<u>Lecture 7: Mill on justice and rights</u>

12. The 'rights'/'justice' objection to utilitarianism
    a. Recall that one of the main objections to act-consequentialism was its violation of deontological constraints: E.g. the case of the sheriff and the innocent scapegoat/the organ-harvesting doctor.
    b. In many such cases, the constraint in question is a particular case of a general principle that one must not *commit injustice*, nor *violate anyone's rights*.
        i. It is *unjust* to punish the innocent man, even if doing so would promote overall utility.
        ii. The innocent man *has a right* not to be imprisoned or executed for a crime he did not commit.
    c. Thus, one common objection to utilitarianism is that it cannot account for the importance of justice, nor of rights.
    d. Mill gives his answer to this charge in Chapter 5 of *Utilitarianism*.
13. Mill's account
    a. Mill seeks to accommodate within utilitarianism, rather than to 'debunk', intuitions regarding justice and injustice.
    b. Mill's initial list of things included in the common idea of justice/injustice
        i. Depriving someone of his *legal* rights.

1. "It is mostly considered unjust to deprive any one of his personal liberty, his property, or any other thing which belongs to him by law."
2. E.g. It is unjust to seize someone's property, even in order to give it to the poor (note the *prima facie* tension with utilitarianism here). It is unjust to imprison someone (to deprive him of his personal liberty), *unless* he has forfeited his legal right to liberty e.g. by committing a serious crime.

ii. Depriving someone of his *moral* rights.
1. We don't want to recognise *only* legal rights, because laws can be bad or incomplete. E.g. It is unjust to keep slaves, even when and where the law does not recognise a universal right to freedom.
2. "We may say, therefore, that a second case of injustice consists in taking or withholding from any person that to which he has a *moral right*" (emphasis in original).

iii. Failure of each person to get what he deserves, either good or bad.
1. "It is universally considered just that each person should obtain that (whether good or evil) which he deserves; and unjust that he should obtain a good, or be made to undergo an evil, which he does not deserve."
2. Sometimes considered to be the whole of what counts as "justice" (cf. e.g. Ross on "prima facie duties of justice"). Mill is using the word in a broader sense, but recognises that the narrower sense is also common.
3. E.g. It's unjust if a virtuous person suffers from persecution and disease; it's unjust if criminals go unpunished.

iv. Breaking faith
1. "It is confessedly unjust to *break faith* with anyone: to violate an engagement, whether express or implied, or to disappoint expectations raised by our own conduct, at least if we have raised those expectations knowingly and voluntarily." (emphasis in original)
2. Obvious example: breaking a promise. This is an *express* engagement.
3. Example of an implied engagement: Suppose that you have been seriously "dating" for several months, in a culture (like ours) in which such relationships are normally expected to be exclusive. You may not have *explicitly promised* to forsake all others (as in a traditional marriage vow). But, unless you've explicitly raised the topic of open relationships, you do have an *implied* engagement not to sleep with other people during the course of this relationship: you have "knowingly and voluntarily" raised an expectation of exclusivity in your partner.

v. Partiality (meaning: being influenced by something that one is not supposed to allow to influence one in this matter)

1. "It is, by universal admission, inconsistent with justice to be *partial*; to show favour or preference to one person over another, in matters to which favour and preference do not properly apply."
2. We need to be a bit careful here, because *sometimes* some significant degree of partiality seems to be actively morally good. "A person would be more likely to be blamed than applauded for giving his family or friends no superiority in good offices over strangers, *when he could do so without violating any other duty*" (emphasis added).
3. Sometimes partiality is bad, but only because it is inconsistent with other, already established, considerations of justice/rights. E.g. "A tribunal, for example, must be impartial, because it is bound to award, without regard to any other consideration, a disputed object to the one of the two parties who has the right to it."
4. Still, it is worth highlighting impartiality as a separate requirement of justice, even if it is strictly speaking redundant. "Impartiality, in short, may be said to mean, being exclusively influenced by the considerations which it is supposed ought to influence the particular case in hand."
5. E.g. It's unjust to favour one's friend over strangers, if one is in the position of deciding who is to get some important job.

vi. 'Equality' (although the extension of this is widely disputed)
    1. Most people regard some sort of 'equality' as being part of justice. But exactly what a principle of 'equality' is supposed to entail varies wildly from one thinker to another.
    2. Example: One important aspect of 'equality' is: equal legal protection and enforcement of rights. But it is consistent with this sense of 'equality' to think that people should have very different rights in the first place, e.g. that slaves have no right to freedom or personal property. "The justice of giving equal protection to the rights of all, is maintained by those who support the most outrageous inequality in the rights themselves."
    3. Also, which inequalities one considers *unjust* depends on which one considers not to be *expedient*. E.g., if one thinks that government is necessary for the good of all, one will not complain of injustice on the grounds of inequality if some people are made into magistrates, and granted powers that are not given to the populace at large. For another example: If one thinks that because of incentive effects, it is necessary to allow those who are more productive to keep more than an equal share of collective production, then one will not consider inequalities in material possessions unjust.
        a. On this line of thought, the inequalities that amount to injustices seem to be those that are *not* justified by some consideration of expediency. (Note that in that case, one cannot wheel in principles of equality to argue against some

measure that is acknowledged to be expedient, i.e. this cannot be a case of the supposed clash between 'expediency' and 'justice' that Mill notes at the start of chapter 5.)

c. On the appropriateness of punishment of some kind
    i. For some of these types injustice, *legal punishment* is appropriate.
    ii. In other cases it is not. For instance, it would be a bad idea to try to punish all instances of promise-breaking (even between friends) by legally sanctioned means: that would be too intrusive, and too costly.
    iii. But even where *legal* punishment is inappropriate, it still seems that punishment *of some kind* is appropriate:
        1. Punishment by public and private disapproval;
        2. Punishment by one's own conscience.

d. A preliminary suggestion
    i. Utilitarianism supplies an obvious test for when punishment of any given kind is 'appropriate': whenever such punishment would, overall, increase the balance of total pleasure over total pain.
    ii. Note that there are *costs* associated with punishments: most obviously, punishment involves inflicting suffering on the punishee. So the benefits of punishment (primarily, the incentive effects that punishment creates) have to be large enough to more than offset these costs, before a utilitarian will recommend the punishment.
    iii. This leads to a preliminary suggestion for how a utilitarian might accommodate considerations of justice and rights: There has been an injustice or a rights-violation any time punishment of some kind would, overall, increase utility.

e. Mill rejects this suggestion, on the grounds that it fails to distinguish between injustices and rights-violations on the one hand, and moral wrongs more generally on the other.
    i. Mill accepts the 'preliminary suggestion' as a utilitarian account of moral *wrong* (as discussed in lecture 4). But a different account is required to single out what is special about *injustice/rights-violation*.

f. Interlude: The distinction between 'perfect' and 'imperfect' duties
    i. Many theorists think that duties can be divided into two kinds: 'perfect' and 'imperfect.'
    ii. Perfect duties are duties that require or forbid *particular* actions. E.g.:
        1. I have a perfect duty to turn up and deliver a lecture on Mill at Exam Schools at 12pm on Fridays.
        2. You have a perfect duty to pay your rent.
        3. We all have a perfect duty not to tell lies (at least under normal circumstances).
    iii. Imperfect duties allow the agent more latitude in precisely how they satisfy the duty. E.g.
        1. A duty to help the needy leaves it open exactly whom, when and where you help.

a. Someone who never does anything to help the needy violates this duty.
b. But you don't necessarily violate the duty by passing by a homeless person on the street. You might meet your duty in some other way than by helping this particular person, e.g. by donating to support homeless charities, or third world health interventions.

2. Some people think that we have a duty to develop our talents. (This would be a duty owed *to ourselves*. Other people think that there are no duties to oneself.) If so, this presumably also leaves it open precisely which talents you choose to develop, when and how.
    a. You would be neglecting this duty if you eschewed every intellectual and artist pursuit (doing a degree, playing the guitar, directing a play) and also every line of useful work for which you are particularly suited (e.g. working as a nurse, if you have a talent for caring for others and reassuring those in distress; teaching, if you have a talent for communicating difficult material and enthusing people about it). I.e., you would be neglecting it if you chose to spend your days working on the supermarket checkout and watching TV in the evenings.
    b. But you don't violate this duty simply by not pursuing *all* of those paths. In any case, for the multi-talented, it would be impossible to develop all one's talents – one cannot e.g. pursue careers as a doctor *and* a lawyer *and* a teacher *and* a musician, all at once.
    c. So there is no *particular* talent that this duty requires you to develop, and there is no particular *way* in which it requires you to develop any of the talents you do select for development. You do not violate this duty e.g. by choosing not to take guitar lessons (because you are taking violin lessons instead, or because you are directing a play instead).

g. Mill claims that the perfect/imperfect distinction lines up with the distinction between violations of justice or rights on the one hand, and other types of moral wrong on the other.
    i. It does, at least at first sight, seem fairly plausible that violations of rights line up with perfect duties.
    ii. Perfect duties and correlative rights
        1. My perfect duty to deliver this lecture lines up with your *right* to hear a lecture from me; I would do you an *injustice* if I failed to turn up (at least without good reason).
        2. Your perfect duty to pay your rent lines up with your landlord's *right* to receive your rent payments in return for housing you.

3. Our perfect duty not to tell lies (under normal circumstances) lines up with our hearers' *rights* not to be deliberately deceived (under normal circumstances).

iii. Imperfect duties do not correlate with rights?
   1. I have a duty to help the needy. But no particular person (e.g. that particular homeless person) has a *right* to my help, precisely because I am free to choose how I fulfil the duty.

h. Anyway: Mill's account of injustice is therefore: I commit an injustice any time I perform some act X such that it would increase utility if my act X were punished (whether by legal, other external or internal means), *and in addition* X is a violation of a perfect, rather than an imperfect, duty.
   i. "Now it is known that ethical writers divide moral duties into two classes, denoted by the ill-chosen expressions, duties of perfect and of imperfect obligation; the latter being those in which, though the act is obligatory, the particular occasions of performing it are left to our choice; as in the case of charity or beneficience, which we are indeed bound to practise, but not towards any definite person, nor at any prescribed time. In the more precise language of philosophic jurists, duties of perfect obligation are those duties in virtue of which a correlative right resides in some other person or persons; duties of imperfect obligation are those moral obligations which do not give birth to any right. I think it will be found that this distinction exactly coincides with that which exists between justice and the other obligations of morality." (U 5.15)
   ii. How this fits our key examples:
      1. The sheriff's duty not to hang innocent Joe Bloggs, if he has any such duty, is a perfect duty. This is therefore (on Mill's account) a matter of justice and rights – presumably, justice towards and rights held by Joe.
      2. The doctor's duty not to harvest organs from the healthy blood-test patient, if he has any such duty, is also a perfect duty, this time owed to the healthy patient in question.

i. Mill anticipates the objection that the notions of rights and justice are more fundamental than his account allows: that they are not simply derivative from considerations about the utility of punishment, but are altogether independent of issues of utility.
   i. Mill's reply, part 1: On that alternative view, there is no possible explanation of *why* the requirements of justice and rights-respecting have the content they do, rather than e.g. the opposite content. (Why not think there is a universal right to the job of Prime Minister, or to welfare benefits of £50k/year; or that justice requires that one's salary is proportional to the position of the first letter of one's mother's Christian name in the alphabet?) In contrast, an account with a foundational utilitarianism in the background can explain this.)

ii. Mill's reply, part 2: One shouldn't resist the utilitarian account on the grounds that it doesn't afford considerations of justice enough *importance*: fundamentality is not the same thing as importance.
    1. "While I dispute the pretensions of any theory which sets up an imaginary standard of justice not grounded on utility, I account the justice which is grounded on utility to be the chief part, and incomparably the most sacred and binding part, of all morality." (U 5.32)

14. Objections to Mill on justice and rights
    a. Objections to Mill's theory of punishment
        i. A *consequentialist* theory of punishment seeks to justify punishment entirely in terms of the consequences of punishing.
            1. This can include consequences of very different types: rehabilitation (of the offender), incapacitation (of the offender, for the protection of society), deterrence (to other would-be criminals).
            2. What it *cannot* (easily?) include is considerations of *retribution*.
                a. Caveat: Moore's theory of goodness. Recall (from lecture 2) that Moore holds that, despite the fact that suffering is bad, a state of affairs that includes both a crime and additional suffering (inflicted on the criminal, by way of punishment) can well be better than a state of affairs including the crime but no corresponding punishment. It is good in itself, according to Moore, if people get what they deserve.
                b. This is how a *consequentialist* account can accommodate retribution. But a *utilitarian* account (i.e. consequentialism + utilitarian theory of the good), like Mill's, cannot accommodate retribution.
        ii. Many non-utilitarians think that for this reason, a utilitarian theory of punishment altogether misses the issue of whether and when punishment is *appropriate*.
            1. Example: Suppose (not entirely implausibly) that it would generally serve as a stronger deterrent if society made a practice of imprisoning, not murderers themselves, but *their mothers*. Should we then adopt that practice? Would it be *appropriate/fitting* to 'punish' the mothers of murderers, or merely expedient to do so?
    b. Objection to Mill's claim that the perfect-imperfect distinction lines up with the justice-rights/other distinction
        i. We saw above that several of our standard examples do seem to fit Mill's account: (on the justice-rights side) our duty not to tell lies, the sheriff's duty not to hang Joe, the doctor's duty not to harvest organs, and (on the non-justice-rights side) the duty to help the needy.
        ii. But other examples *don't* seem to fit.
        iii. Consider e.g. the (putative) duty to develop one's talents. This was supposed to be a duty *to myself*. But presumably, if I have a duty *to myself* to develop (some of) my talents, then I also have a *right* – one that I can claim *against*

myself – to have my talents developed. This is a case in which despite the fact that the duty does not require or forbid any *particular action*, nevertheless there *is* a particular person to whom the duty is owed, and hence there can be a correlative right.

1. Mill might avoid this particular example by denying that there are any duties to oneself.

iv. But other examples, not involving duties to oneself, can be used to make the same point: e.g. I arguably have a duty to develop my *children's* talents, before they reach an age at which this becomes their own business.

v. A possible reply: Mill might just ditch the thesis that the perfect/imperfect distinction lines up with the justice-rights/other distinction, and say instead that the justice-rights/other distinction lines up with whether or not the duty is owed *to any particular person*.

1. In fact he *does* say this:
    a. "In our survey of the various popular acceptations of justice, the term appeared generally to involve the idea of personal right – a claim on the part of one or more individuals, like that which the law gives when it confers a proprietary or other legal right. Whether the injustice consists in depriving a person of a possession, or in breaking faith with him, or in treating him worse than he deserves, or worse than other people who have no greater claims, in each case the supposition implies two things – a wrong done, *and some assignable person who is wronged*.... It seems to me that this feature in the case – a right in some person, correlative to the moral obligation – constitutes the specific difference between justice, and generosity or beneficience." (U 5.15, emphasis added)
2. This quote immediately follows the previous one, in which Mill asserted that the perfect/imperfect distinction is key. Mill does not seem to notice that his two suggestions are *different* and *inequivalent*, as e.g. the example of a duty to develop one's children's talents shows.

c. Objection to Mill's claim that the justice-rights/other distinction lines up with the "wrong done to some assignable person"/other distinction
    i. Some duties are owed to specific persons, but yet don't seem to be matters of rights, nor of justice. E.g. if C has promised to attend D's (large, informal) party, C has a duty to do so, and it is a duty owed to a particular person (namely, D). But C is not committing an *injustice*, nor violating any of the host's *rights*, if she does not attend.
d. Objection to Mill's lumping together justice and rights
    i. Many people think there are important differences between these notions: that neither includes the other.
    ii. Example of a rights-violation that is (arguably) not a case of *injustice*: rape. (This is – arguably – *viciousness* rather than *injustice*.)

i. Example of an injustice that is not a case of rights-violation: arguably, if A nurses B through twenty years of old age, it would be *unjust* for B not to leave anything to A in her will; but it is not that B has a *right* to some of A's estate.

iii. These examples suggest that *no* theory that attempts to draw the justice/nonjustice and the rights/non-rights distinction using a *single* criterion can be correct.

iv. A possible reply: Mill might hold that his account was intended to distinguish between moral wrongs that are matters *either of rights or of justice* on the one hand, and those that are matters *neither of rights nor of justice* on the other.

e. Objection to Mill's claim that his theory holds only that justice isn't *fundamental*, not that it's insufficiently *important*

i. Mill's theory still recommends 'violating rights' in very purified cases. (The Sheriff, again – this is just the point we made back in lecture 3.)

<u>Lecture 8: Two-level and 'global' consequentialism</u>

15. The 'paradox of hedonism'

a. Recall that the hedonist believes that what makes his life go well is for it to contain as much pleasure as possible (or: the greatest possible balance of pleasure over pain).

b. One might therefore assume that, if we set aside other-regarding/moral considerations, the hedonist would recommend *aiming* exclusively at pleasure.

i. One can of course have *instrumental* aims other than for pleasure. For example, there is nothing irrational about aiming *to meet up with my friend*. But according to the hedonist (apparently?), all such instrumental aims that one adopts should ultimately be in the service of the single aim of pleasure. (E.g., I aim to catch the 6pm train to Cardiff, because I aim to get to Cardiff in time to have a cup of tea with my mum, because I think that will be pleasurable (or possibly: because I aim to strengthen my relationship with my mum, which in turn is because I think *that* will lead to more pleasure further down the line – for me, for her, or both).

c. This very natural assumption is, however, false. For any X, it is an *empirical question* whether *humans aiming at X* is part of the most effective means to bringing X about. And in the case of (X=) pleasure: it seems to be the case, as a contingent matter of human psychology, that *consciously and deliberately pursuing pleasure* is often counterproductive: i.e., that the most effective means to pleasure often require that one not hold that aim in mind.

d. An example to illustrate the possibility-in-principle that *consciously and deliberately pursuing* an aim might not be the most effective way of *achieving* that aim: Consider a tennis player whose ultimate aim is to win a Grand Slam. He might well find that obsessing about this aim while he plays is counterproductive: it raises his stress levels to an unhelpful degree, and it distracts him from exercising his natural intuitive feel for the game in order to play well. His coach might well advise him to

try not to think about winning while he plays, and allow himself just to get absorbed in the game, *in order to increase his chances of winning*.

e. Examples to suggest that the same point does arise in the case of the ultimate aim of one's own pleasure:
    i. One of the paradigmatic "higher pleasures" is the pleasure of intellectual enquiry and adventure (hopefully, the pleasure that you get from your studies, in their best moments!). But one does not get this pleasure if one is constantly *thinking about* the fact that one is studying in order to achieve this pleasure: one needs to some extent to 'lose oneself' in the subject itself, driven by pure intellectual curiosity.
    ii. Many people would enjoy a game of cricket or chess less if their deliberate aim throughout was the pleasure of playing, rather than some attempt at winning. I.e. this issue might apply to *enjoying* a game, as well as (as in the tennis example above) to *winning* the game. [Anecdote from playing games with Tim!]
    iii. One can gain great pleasure from close relationships (whether 'platonic' or romantic). But one would probably not succeed in forming good relationships in the first place if one's sole aim throughout was merely to enjoy the pleasures that will result once one has formed one: part of what it means to have a close friendship is e.g. that to some extent one treats one's friend's projects, *and the aims that they involve adopting,* as one's own. (Parrot-keeping.)
    iv. One can gain great pleasure simply from observing the world if one has a "strong and lively interest in the well-being of prosperous persons and institutions", but not if one is merely trying to *fake* such an interest in order to obtain the pleasure in question. (Cf supporting a successful sports team.)

f. A naive reaction: therefore hedonism is false.
    i. This is too naive.
    ii. Note that the 'paradox' of hedonism isn't really a *paradox*, and that no-one has suggested that hedonism literally *entails a contradiction*.
    iii. But still, there does seem to be something unsettling here: it seems to be the case that, in some sense, the hedonist 'ought not to be a hedonist'.

g. Attempting to make this more precise
    i. There is a threat of a 'reductio' argument against hedonism: Suppose, for the sake of argument, that hedonism is true: that pleasure and only pleasure is good. Plausibly(?), one is rationally required to aim at things one believes to be good. So, if one *believes* that hedonism is true, then one is rationally required to aim at pleasure. But it is a general principle of rationality that, if one is rationally required to aim at X, one is also rationally required to aim at whatever are the best means of achieving X. But we noted above that, as a matter of contingent psychological fact, the best means of achieving pleasure include *not aiming at pleasure.* Therefore, if one is rationally required to aim at pleasure, one is also rationally required to aim *not to aim at pleasure*. But if one *succeeds* in this second aim, then one fails to aim at pleasure. It therefore seems that if hedonism is true, one is condemned to

be irrational: one is irrational if one fails to aim at pleasure in the first place, one is irrational if one aims at pleasure but fails to aim not to aim at pleasure (despite knowing that not-aiming-at-pleasure is the best means of achieving one's established aim), and one is irrational if one obeys all the requirements of rationality so far stated and in addition succeeds in achieving the aims one has thereby adopted.

h. Sidgwick's reply to this worry
    i. It is indeed impossible to hold a desire for pleasure and some of the desires whose adoption is the most effective means to achieving pleasure *simultaneously*.
        1. "[T]hough we could distinguish appetite, as it appears in consciousness, from the desire of the pleasure attending the satisfaction of appetite, there appeared to be no incompatibility between the two. The fact that a glutton is dominated by the desire of the pleasures of eating in no way impedes the development in him of the appetite which is a necessary condition of these pleasures. But when we turn to the pleasures of pursuit, we seem to perceive this incompatibility to a certain extent: a certain subordination of self-regard seems to be necessary in order to obtain full enjoyment." (Methods of Ethics, p.48)
    ii. But all that follows from this is that really adopting non-hedonistic aims psychologically requires one to *forget* one's hedonism, *temporarily*. But hedonism can still play a role, by being consciously held as an aim at other times, and by being the background consideration that guides one's choice of which other aims to adopt.
        1. "[I]n the ordinary condition of our activity the incompatibility [between desires for pleasure and the non-hedonistic desires that cause pleasure] is only momentary, and does not prevent a real harmony from being attained by a sort of alternating rhythm of the two impulses in consciousness." (ME, p.136)
        2. And this "alternating rhythm" is perfectly psychologically possible: "[I]t is an experience only too common among men, in whatever pursuit they may be engaged, that they let the original object and goal of their efforts pass out of view, and come to regard the means to this end as ends in themselves: so that they at last even sacrifice the original end to the attainment of what is only secondarily and derivatively desirable. And if it be thus easy and common to forget the end in the means overmuch, there seems no reason why it should be difficult to do it to the extent that Rational Egoism prescribes: and, in fact, it seems to be continually done by ordinary persons in the case of amusements and pastimes of all kinds." (Sidgwick, ME, p.137)
    iii. How this might play out in the context of our examples
        1. The tennis player: The player doesn't focus on winning *while he is playing*. But in discussions with his coach in between games, he may

well revert to the ultimate criterion of what is likely to maximise his chances of winning, in order to strategise about what the precise attitude is that he should adopt while he is on the court, and in order to think about how to get himself into that state of mind for his games.

2. The game of chess: The player doesn't *normally* focus on getting pleasure from the game *while he is playing*. But between games, he is likely to think about whether his chess-playing habit is making him happier or less happy, and might well stop playing if he decides the latter.
3. The sports spectator: The spectator doesn't focus on the fact that pleasure is the point of the whole enterprise for him while he is watching the match. But, from time to time, he does reflect on whether being a sports fan is succeeding in making his life more pleasurable, and he might decide to stop going to matches if he found that he wasn't enjoying them.
4. The friend: I don't focus on the fact that I'm in it for the pleasures of friendship *while I'm thinking through with my friend how to build a parrot cage*. But, from time to time, I will reflect on the degree to which it is optimal to get involved with a friend's personal projects, and it may very well be considerations of pleasure (my own *and my friend's*) that govern those reflections.

16. The 'self-defeatingness' objection to (act-)consequentialism
    a. Similarly: many theorists have noted that *deliberately trying* to bring about as much utility as possible on an act-by-act basis may have the unfortunate consequence of bringing about less utility than one would have brought about if one had *adopted* some alternative, apparently non-consequentialist, decision procedure. There are several reasons for thinking that this might happen.
    b. The calculation-time worry
        i. Someone who tried to carry out a full consequentialist calculation every time (s)he made any decision would spend most of her life calculating, rather than getting on with acting.
            1. First example: This may lead to missing the crucial opportunities for action: Suppose you are going for a run by the river, and see a child drowning. If you stop to think through all the possible long-run consequences of jumping in to save her, then by the time you have made the decision that it is best to jump, it will be too late.
            2. Second example: In many cases it is predictably just a waste of time: If you attempt a full consequentialist calculation for every trivial decision, such as whether to go left or right round the block on the way to your lecture, your mind will be exhausted by millions of pointless calculations – itself a source of stress and unhappiness, and one that distracts you from doing more useful things with your mental time and energy.
    c. The worry of personal bias

i. Predictions of the overall balance of good over bad that is likely to result from one's actions require judgment calls at every turn – regarding both how likely the various possible consequences are, and how good or bad they are.
    1. Example: You find a wallet on the street, and are deliberating about whether to return it to its owner or not. In the heat of the moment, It's all too easy to overestimate the probability that the owner is rich enough that losing his wallet won't really matter to him, and to ignore or underestimate the stress and inconvenience of having to cancel all one's cards, when one's own interests immediately depend on it.
    2. Example 2: You are an officer in charge of a charity raffle, carrying a prize of £10,000. You happen to know that ticket 5348 belongs to your friend, who desperately needs the money. You must decide whether to carry out the raffle draw fairly, or to announce ticket 5348 as the winner. In this situation, if you were to attempt an explicit consequentialist calculation (rather than e.g. adhering to a general principle of integrity in adhering to the role you have accepted), it would be easy to underestimate the probability of discovery, and of the knock-on damage via ruining the reputation of the charity, if it seems *reasonably* certain that you could direct the money to your friend without anyone ever being any the wiser.
    3. Another example: Your promise to meet Aunt Nellie for dinner, again

d. The worry of alienation
    i. Example: Your friend is ill in hospital. You must decide whether to go and visit him, or whether to spend the evening volunteering at the local soup kitchen instead. After weighing up all the possible ways that each course of action might lead to good, including estimates of their probabilities and relative levels of goodness, you decide that all things considered, it is better to visit your friend. You explain all this to him when he comments that it's nice of you to come and visit.
    ii. Intuition: There's *something wrong* with visiting your friend only for this reason, and only when reasons of this character recommend doing so. Your friend would justified in feeling aggrieved when you tell him that these were your reasons.
    iii. Railton's diagnosis of what's wrong (in his article 'Consequentialism, alienation, and the demands of morality'):
        1. In the above sort of case, there is an 'estrangement' between the agent's *affections* and her *rational, deliberative self*.
            a. The affections essentially involve partiality. But the rational, deliberative self is rigidly following a completely impartial morality.
        2. This alienation might be bad because: "it may be the basis of... a sense of loneliness or emptiness – or the loss of certain things of

value – such as a sense of belonging or the pleasures of spontaneity. Moreover, their alienation may ... make certain valuable kinds of relationships impossible."

3. This line of thought suggests that in this example too, *deliberating on the basis of consequentialism* might itself have bad consequences.

e. Conclusion: For all these reasons, even a consequentialist should not, by his own lights, desire that people always (or even often) make decisions by carrying out explicit consequentialist calculations (just as a hedonist should not, by his own lights, desire that people always (or even often) make decisions by carrying out explicit hedonistic reasoning).

f. Does this show that consequentialism is self-defeating – that "the consequentialist should not be a consequentialist"?

i. Williams: Yes. "Utilitarianism's fate is to usher itself from the scene" ("A critique of utilitarianism", in Smart & Williams, p.134)

ii. But the analogue of Sidgwick's reaction to the paradox of hedonism is again a very different reaction: the above considerations do show that a consequentialist should try to inculcate (both in herself, and in others) a tendency to deliberate in non-consequentialist ways, but they don't show that there is *no* remaining role for consequentialist ways of thinking.

17. Hare's version of this view (in his "Moral thinking", chs. 2 and 3)

a. Hare advocates a "two-level" view of moral thinking: an "intuitive" level, involving such principles as 'don't steal', 'don't lie' etc., and a "critical" level, involving evaluation in terms of likely consequences.

b. If there were any "archangels" – ideal beings not subject to human limitations – then the best way for *them* to proceed would be to reason on the critical level all the time.

i. An archangel (by definition):

1. calculates at infinite speed, so the 'calculation time' worry does not apply to him.
2. is entirely free from personal bias, so that worry does not apply to him.
3. feels genuine affection for strangers just as much as for friends, so the "alienation" worry does not apply to him.

c. On the other hand, if there are any "proles" – extremely stupid beings, who would get the wrong answer any and every time they tried to do any consequentialist calculation – then the best way for *them* to proceed is to reason on the intuitive level all the time.

i. This doesn't help much if *everyone* is a prole, since in that case there's nothing to ensure that the intuitive principles are good ones. But if someone in the community is able to do such calculations at least sometimes, then he can take charge of supplying the proles with sensible principles. (Cf. the raising of children.)

d. Most or all of us, though, are neither archangels nor proles, but *something in between*. It is therefore optimal for us to follow a *combination* of the "archangel" and "prole" strategies:
    i. Much of the time, we should simply follow the intuitive principles.
    ii. But when principles conflict, or in the relatively rare situations in which it seems clear that the recommendations of the principles are silly, we can shift to critical thinking to work out how to resolve the conflict/issue.
        1. Examples:
            a. Easy case: Lying to the murderer at the door
            b. Harder case: State-sponsored (and *morally* motivated) assassinations – consider, e.g., a plot to assassinate Hitler prior to World War II. *If* statesmen in 1938 had foreseen what Hitler would go on to do, arguably they would have been justified in plotting to have him assassinated, despite the fact that it is (probably) good if they sign up to a *general* rule against assassinating foreign leaders. And *if they wouldn't, the reason for that must in turn be in terms of even more serious knock-on consequences.*
    iii. And in any case, at some times we should engage in critical thinking in order to decide whether the intuitive principles that are currently "programmed" into us are the right ones.
        1. Examples:
            a. One's upbringing might have led one to have a habit visibly of expressing disgust when someone breaks a rule of table etiquette, but critical reflection may later convince one that this disposition is not for the best.
            b. On calculation time: At some point(s) in one's life, one should critically evaluate the amount of consequentialist calculation one does on a day-to-day basis, and whether it is damaging and/or cost-ineffective.
            c. On personal bias: One should think carefully about the sorts of situations in which personal bias is a serious worry, and when it can safely be ignored. E.g., perhaps I know that I am susceptible to personal bias when trying to think consequentialistically about whether to keep unwanted social commitments, so I am particularly careful to "just stick to the rules" in my day-to-day practice on those particular matters.
            d. On alienation: One should think carefully about what are the appropriate *limits* to simply following one's emotions. E.g., many people think that one should give a significant proportion of one's money to charity, *despite* the fact that one's immediate desires favour much more partiality to self, family and friends. If I think this, then when my child asks for an expensive Christmas present and I have to choose

between buying him that and keeping up my regular donation to the Against Malaria Foundation, it doesn't seem at all inappropriate to invoke the consequentialist explanation of why I am not going to buy him the latest video-game gizmo.

e. The fact that we do, and should, have intuitive and emotion-laden reactions against actions that are forbidden by the "intuitive" rules explains the sense in which it is appropriate to feel, not guilt as such, but *compunction*, when one breaks an intuitive rule on the basis of a critical judgment that on this occasion breaking the rule is for the best.
    i. Example: Williams' case of Jim and the Indians
        1. The case: "Jim finds himself in the square of a small South American town. Tied up against the wall are a row of twenty Indians, most terrified, a few defiant, in front of them several armed men in uniform. A heavy man in a sweat-stained khaki shirt turns out to be the captain in charge and, after a good deal of questioning of Jim which establishes that he got there by accident while on a botanical expedition, explains that the Indians are a random group of the inhabitants who, after recent acts of protest against the government, are just about to be killed to remind other possible protesters of the advantages of not protesting. However, since Jim is an honoured visitor from another land, the captain is happy to offer him a guest's privilege of killing one of the Indians himself. If Jim accepts, then as a special mark of the occasion, the other Indians will be let off. Of course, if Jim refuses, then there is no special occasion, and Pedro here will do what he was about to do when Jim arrived, and kill them all. Jim, with some desperate recollection of schoolboy fiction, wonders whether if he got hold of a gun, he could hold the captain, Pedro and the rest o the soldiers to threat, but it is quite clear from the set-up that nothing of that kind is going to work: any attempt at that sort of thing will mean that all the Indians will be killed, and himself. The men against the wall, and the other villagers, understand the situation, and are obviously begging him to accept. What should he do?" (Williams, 'A critique of utilitarianism', in Smart & Williams, *Utilitarianism: For and against*, pp.98-9)
        2. Williams raises this case as an objection to utilitarianism, the point apparently being that although utilitarianism probably gives the right answer to the question "Should Jim shoot or not?" in this case, it fails to capture the fact that Jim should make this decision *very reluctantly*, and that he should *feel very bad* about shooting.
        3. Hare's response: Unlike a simple act-consequentialism, the two-level view *does* capture the sense in which Jim "should feel bad": the optimal pattern of reactions for Jim to have is one including a strong intuitive aversion to killing, and if he does, then it is

inevitable that he will feel a strong sense of compunction if he shoots the first Indian.

ii. Hare's example: Lying to Czech immigration about the purpose of his visit

18. Williams against two-level consequentialism
    a. Hare's two-level consequentialism recommends thinking on the "intuitive" and the "critical" levels in precisely the sort of "alternating rhythm" that Sidgwick recommended in the context of the paradox of hedonism.
    b. Williams thinks that it is psychologically impossible to carry out this procedure with the intended results:
        i. "If an agent has no more than the thought that it is instrumentally useful for him to think that a certain value [such as the value of friendship] is not merely instrumentally useful, the structure... will come to no more than a pretence, and for that reason it will be unstable under reflection. This is what happens to indirect Utilitarianism, the kind of theory that recommends, on strictly Utilitarian principles, rules or disposition which will lead us to choose certain actions that, in themselves, would not be chosen by a Utilitarian. The trouble with this is that an agent who needs to reflect on a situation in which he is disposed to do such an action has no thought to fall back on except that it is Utilitarianly valuable that he should have this disposition, and this leaves no content to the disposition: he has no thoughts with which to counter the consideration that some alternative action in this situation is the one that has the best Utilitarian consequences."
    c. Hare is unimpressed by this criticism: "I can only reply by asking whether those who raise this objection have ever faced such situations [of conflict between normally-effective moral 'rules of thumb']. I do my own moral thinking in the way described in this book (not like an archangel, for I am not one, nor like a prole, but doing my best to employ critical and intuitive thinking as appropriate)." (Hare, *Moral Thinking*, p.52)
19. Railton
    a. Railton also proposes a sort of "two-level" view, but formulates it differently from Hare.
    b. Response to the paradox of hedonism: distinguish between "subjective" and "objective" hedonism:
        i. Subjective hedonism: "One should whenever possible attempt to determine which act seems most likely to contribute optimally to one's happiness, and behave accordingly."
        ii. Objective hedonism: "One should follow that course of action which would in fact most contribute to one's happiness, even when this would involve *not* adopting the hedonistic point of view in action."
        iii. 'Sophisticated hedonist': one whose aim is to obey objective hedonism, and who realises that (as the 'paradox' of hedonism points out) this aim requires *not* obeying subjective hedonism.
    c. Response to the "self-defeatingness" objections to consequentialism: distinguish between "subjective" and "objective" consequentialism

i. Subjective consequentialism: One should whenever possible attempt to determine which act seems most likely to contribute optimally to overall good, and behave accordingly. ("Consequentialism as a decision procedure.")
ii. Objective consequentialism: One should follow that course of action which would in fact most contribute to overall good, even when this would involve *not* adopting the consequentialist point of view in deliberation. ("Consequentialism as a criterion of the right.")
iii. 'Sophisticated consequentialist': One whose aim is to obey objective consequentialism, and who realises that (because of the concerns raised in the course of the "self-defeatingness" objection) this aim requires not using consequentialism as a decision procedure.

20. Two-level consequentialism summarised:
    a. The *criterion of right action* is as act-consequentialism says.
    b. The recommended *decision procedure* is: whichever procedure has the property that *adopting it as one's decision procedure* would lead to better consequences overall than adopting any alternative decision procedure.
    c. What the self-defeatingness objection points out is (only) that the optimal decision procedure is not: perform an explicit act-consequentialist calculation for every action (and inaction) that you take.
    d. What the optimal decision procedure *is* likely to be:
        i. Reasonably simple, objectively applicable 'rules of thumb' that experience has shown *generally* to be conducive to the promotion of overall goodness.
21. 'Global' consequentialism
    a. Motivating thought: Why only a criterion of right action and a decision procedure – couldn't we apply the basic consequentialist maxim "maximise goodness" to other things, too?
    b. The global consequentialists' idea is that the basic consequentialist principle of maximising goodness (or realising at least a certain threshold level of goodness (satisficing version), or realising more goodness rather than less (scalar version)) can be applied not only to acts and decision procedures, but to arbitrary propositions.
    c. Examples: preliminary version
        i. Character traits: The best set of character traits is the one that would maximise overall goodness
        ii. Laws: The best set of laws is the one that would maximise overall goodness
        iii. Forms of government: The best form of government is the one that would maximise overall goodness... etc.
    d. Clarification: the set of character traits such that *possessing* those character traits would maximise goodness, or such that *trying to acquire* those traits would maximise goodness, or such that *praising* those traits would maximise goodness, or what?
    e. Reply: These are just answers to distinct questions. Global consequentialism needs to be reformulated to recognise the distinctness:
        i. The best set of character traits *for Alf to possess* is the one whose *possession by Alf* would maximise overall goodness.

22. Suggestions of global consequentialism in Mill's *Utilitarianism*?
    a. "[Utilitarianism] maintains not only that virtue is to be desired, but that it is to be desired disinterestedly, for itself... [T]he mind is not in a right state, not in a state conformable to utility, not in a state most conducive to the general happiness, unless it does love virtue in this manner – as a thing desirable in itself, even although, in the individual instance, it should not produce those other desirable consequences which it tends to produce, and on account of which it is held to be virtue." (U 4.5)
23. Objections to two-level and global consequentialism
    a. Objection-type 1: Objections to the act-consequentialist criterion of right/wrong action.
        i. Note that neither the two-level nor the global consequentialist drops the act-consequentialist's *criterion of right action* (they merely emphasise that the criterion of right action is not all there is to morality). Therefore, any objection to act-consequentialism *that really is an objection to that criterion of right action* remains as an objection to two-level or global consequentialism.
            1. E.g. in the purified sheriff case, two-level and global consequentialists agree that – however a good moral agent is likely to be *feeling*, etc. - the *right action* is to hang the innocent man.
            2. Hare asserts (pp.48-9) offers an explanation of the widespread tendency to insist otherwise as an understandable *mistake* on the part of the anti-utilitarian. (According to Hare, this is a case in which critical thinking ought to kick in, and the good moral agent should override even his most deeply ingrained "rules of thumb" screaming the contrary verdict.)
        ii. Those who are unconvinced by this, and who continue to insist that the sheriff should not hang the innocent man, however, will continue to take this to be an objection to a two-level/global consequentialism, just as much as it is an objection to basic act-consequentialism.
    b. Objection 2: Internal inconsistency?
        i. Suppose that in a given decision situation, the *act* recommended by global consequentialism is distinct from the act that would be selected by the *decision procedure* that is recommended by global consequentialism. Then it is impossible to satisfy all the demands of the theory simultaneously.
        ii. Defenders of two-level/global consequentialism have to hold either that this cannot after all happen (well – can it??), or that the result is not any *problematic* sort of 'inconsistency' if it does happen.