Foundations of Effective Altruism
Frank Arntzenius and Hilary Greaves
Hilary Term 2017
Week 1: Introduction

Logistics

This course has a Weblearn site. Reading lists, together with core readings, will be posted at least a week in advance. Please read the assigned 'core readings' before attending each class.

Here is a provisional schedule of topics for this term:

Week 1: Introduction; Singer's argument
Week 2: Pogge's argument
Week 3: Making a difference
Week 4: Career choice
Week 5: Identified vs statistical lives
Week 6: Conditional obligations
Week 7: Existential risk and population ethics
Week 8: Cluelessness and indirect effects

If there is sufficient interest, we may continue the course into Trinity term.

This course focusses on academic moral philosophy topics that are relevant to the EA movement. If you are unfamiliar with the EA movement itself, a good non-academic introduction is Will MacAskill's *Doing Good Better*.

Background: The effective altruism (EA) movement

1. A (very) brief history of EA
    a. 1972: Peter Singer published 'Famine, Affluence and Morality'
    b. 2007: Holden Karnofsky and Elie Hassenfeld started GiveWell: evaluating charities for effectiveness with a view to advising philanthropic donors (esp. small donors). Most early recommendations were in the field of poor-country health.
    c. Meta-research on other possible recipients for charitable giving
        i. Research on cost-effective philanthropic policy interventions: Open Philanthropy Project (2012), OpenBorders.info (2012)
        ii. Animal welfare interventions: Sentience Politics (2013), Animal Charity Evaluators (2013)
    d. Movement growth activities
        i. Encouraging small donors to give more and to give effectively: Giving What We Can (2009), The Life You Can Save (2013), Charity Science (2013)
        ii. "EA Global" and EAGx conferences (since 2013), Singer's and MacAskill's books (both 2015)[1]
    e. Advice on altruistic career choice: 80,000 hours (2012)

[1] MacAskill, *Doing Good Better*; Singer, *The most good you can do*

f. Raising standards of rationality: Center for Applied Rationality (2012, California)
g. Research on existential risk reduction: Future of Humanity Institute (2005, Oxford), Centre for the Study of Existential Risk (2012, Cambridge), Machine Intelligence Research Institute (2000, California), Future of Life Institute (2014, Boston), Foundational Research Institute (2013, Switzerland)

2. Key aspects of the movement
    a. Effective altruists define the movement as follows: "Effective altruism is about using evidence and careful reasoning to address the question, "How can I do the most to help others with my time or money?", and taking action on the basis of your answer."[2]
    b. Important aspects
        i. Taking action: making altruism a central part of one's life
        ii. 'Doing the most to help others': this covers
            1. Cost-effectiveness[3] - more important than *how much* you give.
            2. Cause-neutrality: not e.g. "what you have a passion for"
        iii. Evidence and careful reasoning: prioritisation is an in-depth process
    c. Some common misconceptions, and replies
        i. "EA is just about global poverty."
        ii. "EA is just utilitarianism." (Or: consequentialism.)
        iii. "EA neglects systemic change".
    d. Perhaps not in the definition, but also important: A focus on outcomes
        i. The relevant thing is how much better off others are as a result of one's action, *compared to the situation in which one had acted differently*. Not the same as one's 'direct impact' (cf. earning to give, meta-work).

The argument from utilitarianism for poverty relief

3. Textbook utilitarianism:
    a. Theory of well-being: Hedonism: Individual well-being consists in happiness/pleasure and absence of unhappiness/pain.
        i. (Rather than preference-satisfaction/objective-list-ticking).
    b. Theory of aggregation: 'Utilitarian aggregation': Overall good is the (impartial, i.e. unweighted) sum of individuals' well-being.
        i. (Rather than e.g. a prioritarian or egalitarian 'value function'.)
    c. Deontology: Maximising consequentialism: An act is right iff it leads to the best available outcome.
        i. (Rather than e.g. satisficing consequentialism, Rossian common-sense intuitionism, Kantian deontology, …)
4. Sketch of a utilitarian argument for poverty relief
    a. Suppose x is worse off than you. Then:
        i. If you gave half your 'excess' wealth to x, the amount by which you would increase x's happiness is greater than the amount by which you would decrease your own happiness. *(Diminishing marginal utility)*

[2] MacAskill, 'The definition of 'effective altruism'', forthcoming.
[3] Ord, 'The moral imperative towards cost-effectiveness'

ii. You are required to act so as to maximise total happiness. *(Utilitarianism)*
Therefore,
iii. You are required to give half your excess wealth to x [if this is your only option aside from doing nothing].

b. Obvious application: Famine relief charities

5. Obvious worry: Demandingness
   a. The argument iterates, until you have given away so much that you yourself are (equally as poor as) the world's poorest. Does morality really require giving away *this* much??
   b. This is one major source of resistance to utilitarianism (it is one aspect of "the demandingness objection to utilitarianism").

Singer's argument in "Famine, affluence and morality"

6. Singer's argument
   a. Two cases
      i. Famine Relief: Children in the Third World are starving. You have a choice as to whether or not to send money to a famine relief charity. Sending $100 would make the difference between life and death for several children.
      ii. Shallow Pond: On your way to a lecture you pass a child in danger of drowning in a shallow pond. You could jump in to save him. If you do, he will be saved, but your expensive suit will be ruined/you will miss the lecture/etc.
   b. Singer's argument
      (P1) "Suffering and death… are bad."
      (P2) "If it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance/anything morally significant, we ought, morally, to do it."
         i. Illustrative example: The shallow pond
         ii. Note that <u>this is *not* an appeal to utilitarianism</u>.
      (P3) By giving substantial sums of money to famine relief charities, we can prevent suffering and death, without sacrificing anything of comparable moral importance/anything morally significant.
         iii. Empirical estimate (2015): You can save a child's life by donating approx. US$3,500 to the Against Malaria Foundation.
      Therefore,
      (C) We ought to give substantial sums to famine relief charities.
7. Objection 1: Where this argument leads is "too demanding"
   a. On demandingness
      i. "Without giving up anything morally significant" (at any rate) is a lot stronger than "without giving up more happiness than you would cause the beneficiary". (Also, there's a lot of room for interpretation/disagreement over what counts as 'morally significant'.) So Singer's conclusion is less demanding than the utilitarian one (but might still be very demanding).
8. Objection 2: Overgeneralisation

a. Insofar as Singer argues for his crucial premise (P2), the argument seems to be by generalisation from a single case:
    i. "An application of this principle would be as follows: if I am walking past a shallow pond and see a child drowning in it, I ought to wade in and pull the child out. This will mean getting my clothes muddy, but this is insignificant."
b. But false principles can have true instances. Couldn't there be some relevant disanalogy between the case of 'Shallow Pond' and that of 'Famine Relief'?
c. Singer seems to be really (just) arguing as follows ("the argument by analogy"):
    i. You are morally required to help in Shallow Pond.
    ii. There's no morally relevant difference between Shallow Pond and Famine Relief.
    iii. Therefore, you are morally required to help in Famine Relief.
d. Unger's defence of Singer: the numerous differences between Shallow Pond and Famine Relief really *are* morally irrelevant
    i. Unger's two methodologies:
        1. Consider, for the suggested difference between Sedan and Envelope, variants of the two cases that remove the difference in question, and note that the intuitions don't change.
        2. Observe that it's just not (common-sensically) plausible that the difference in question is morally relevant.
    ii. Differences that Unger argues are morally irrelevant include: physical proximity, social proximity, informative directness, experiential impact, number of potential benefactors, number of potential beneficiaries, degree of urgency…
e. The conservative response
    i. We grant that there's a puzzle: we don't know what the morally relevant difference between Shallow Pond and Famine Relief is. But clearly there must be *some* morally relevant difference, because the consequences if there isn't are too demanding.

Effective altruism and non-utilitarian moral theories

9. Spoiler: The basic points
    a. Almost all moral theories agree that helping others is at least supererogatory, and many agree that to some extent it is morally required.
    b. Almost all moral theories agree that outcomes are important.
    c. EA does not, and need not, recommend or condone violations of rights or of moral "side constraints".
10. Some other moral theories
    a. Satisficing consequentialism[4]: in any given decision situation, there is some threshold amount of goodness, G, such that you are morally required to perform some act whose [expected] consequence-goodness is at least G. Performing an act that whose outcome is *better* than this required minimum is supererogatory: morally better than the required minimum, but not actually required by morality.

[4] Slote, 'Satisficing consequentialism'.

b. Scalar consequentialism[5]: In any given decision situation, one act is better than another iff its [expected] consequences are better.
c. Consequentialism subject to side-constraints: actions with certain features (e.g. those that involve killing, or breaking promises) are (normally) morally forbidden; use your favourite from maximising/satisficing/scalar consequentialism to issue verdicts on the remaining actions.
d. Schefflerian theory of agent-centred prerogatives[6]: in ranking outcomes in terms of their consequences, an agent is morally permitted to assign somewhat higher weight to her own interests and projects, and those of people near and dear to her, than she does to the interests and projects of strangers. (Then apply maximising 'consequentialism' to the resulting ranking.)
e. Rossian intuitionism[7]: There are several 'prima facie duties' [modern moral philosophers would say: several 'pro tanto moral reasons'] that might count for or against particular actions. These include duties [/reasons] of fidelity (keeping promises), reparation (making up for past wrongs), justice (making sure that what people get is in proportion to what they deserve)… but also 'duties of beneficence'.

11. Virtue ethics (see e.g. https://plato.stanford.edu/entries/ethics-virtue/)
    a. Virtue ethicists approach the question of what to do via the (to them more central) question of *what kind of person to be*.
    b. What kind of person to be:
        i. A *virtuous* person, i.e. one who possesses and exercises the virtuous character traits to a high degree.
        ii. 'Virtuous character traits' are typically taken to include: courage, honesty, justice, *generosity*, …
    c. At least arguably, the appropriate form of generosity includes a concern for effectiveness, as well as a concern to be doing something/making major sacrifices for the benefit of others.[8]
12. "What about Kant?" (see e.g. https://plato.stanford.edu/entries/kant-moral/)
    a. Kantian moral philosophy might be thought to be the most hostile to EA, and to philanthropy in general.
    b. But even Kant recognises a duty to help others:
        i. "We have a duty to be charitably helpful where we can" (*Groundwork*, chapter 1).
        ii. In more detail: Kant argues that a maxim of never helping others fails his universalizability test, since we could not 'will' a world in which everyone operated on a maxim of never helping others.[9]

---

[5] Norcross, 'The scalar approach to utilitarianism'.
[6] Scheffler, *The rejection of consequentialism.*
[7] Ross, *The right and the good.*
[8] Chappell, 'Overriding virtue', in preparation.
[9] "A fourth man, for whom things are going well, sees that others (whom he could help) have to struggle with great hardships, and he thinks to himself:

> What concern of mine is it? Let each one be as happy as heaven wills, or as he can make himself; I won't take anything from him or even envy him; but I have no desire to contribute to his welfare or help him in time of need.

iii. Kant holds that this is an "imperfect" duty: there is latitude about precisely how one fulfils it (thus there is no *particular* helping action one is required to do, and no particular person one is required to help, by this duty).

c. It's not clear what Kant would say about effectiveness or cause-neutrality (since he doesn't explicitly discuss the topics…)

13. "Is effective altruism trivial then?"

a. Possibly, when construed as a 'moral philosophy'. But (1) it is highly nontrivial as a social movement, and (2) it helpfully highlights certain other research questions (about *how* to do the most good) as particularly important, as well as (3) pointing out as errors certain common ideas that might otherwise have slipped under the radar.

If such a way of thinking were a universal law of nature, the human race could certainly survive—and no doubt that state of humanity would be better than one where everyone chatters about sympathy and benevolence and exerts himself occasionally to practice them, while also taking every chance he can to cheat, and to betray or otherwise violate people's rights. But although it is possible that that maxim should be a universal law of nature, it is impossible to *will* that it do so. For a will that brought *that* about would conflict with itself, since instances can often arise in which the person in question would need the love and sympathy of others, and he would have no hope of getting the help he desires, being robbed of it by this law of nature springing from his own will." Kant, *Groundwork of the metaphysics of morals*, chapter 2

Foundations of Effective Altruism
Frank Arntzenius and Hilary Greaves
Hilary Term 2017
Week 2: Pogge's argument

1. Background: Some key distinctions
    a. Positive vs negative duties
        i. "Positive" duties are duties *to do* particular things (e.g., the duty to rescue the drowning child/to help the needy). "Negative" duties are duties to refrain from doing particular things (e.g., the duty to refrain from stealing).
            1. Presupposes a distinction between action and inaction.
        ii. Generally speaking, in common-sense morality, negative duties tend to be 'more stringent' than positive ones - one commits *more of a wrong* by violating them.
            1. If we hold fixed how much is at stake (in consequentialist terms).
        iii. Complication: sometimes one acquires a ('derivative') positive duty *as a result of* having previously violated a negative one. These are more stringent, like the original negative duty.
            1. E.g. paying back something one has stolen.
    b. Duties of beneficence vs duties of justice (and others)[10]
        i. Beneficence: The mere fact that someone else would benefit from your doing X is *a* reason for you to do X.
            1. Examples: X = donating a meal to a food bank, visiting strangers in hospital.
            2. NB: It does not necessarily follow that you *ought to do* X, since reasons can in general be overridden by other, competing reasons.
        ii. Justice[11]
            1. Examples: refraining from stealing, keeping your contracts.
            2. First demarcation criterion: duties of justice are those that are *owed to particular other persons* (as opposed to: duties owed to no-one, duties owed to oneself).
            3. Second demarcation criterion: If X is a duty of justice, it is permissible for (some) third parties to *use force to* (i) compel agents to comply with X and (ii) punish non-compliance.
            4. Examples
                a. The above examples of duties of justice meet both these criteria, while the above examples of duties of beneficence meet neither.
                b. Things that meet one criterion but not the other
                    i. Visiting your mother.
                    ii. Perhaps: paying your taxes??
        iii. Generally speaking, according to common-sense morality, duties of justice tend to be more stringent than duties of beneficence.
    c. A sub-class of duties of justice: duties of *reparation*

[10]W. D. Ross, "The right and the good", section 2 ("what makes right acts right?", especially the numbered list of 6 types of duty. Available online at http://www.ditext.com/ross/right2.html
[11]Broome, Climate Matters, chapter 4, esp. pp.50-54 ("goodness versus justice")

i. Reparation: You have previously violated some duty of justice, and as a result of that, you are now under a corresponding duty of reparation.

2. Relevance of these classifications to the debate over the moral motivations for 'helping' the global poor
    a. The key question for 'EA' purposes: are our duties to 'help' the global poor duties of (mere) beneficence, or (also) duties of reparation or otherwise of justice?
    b. Singer's discussion is consistent with them being duties of mere beneficence. Many commentators (incl. Pogge) hold that they are duties of justice. (Correspondingly, these commentators dislike the terminology of 'helping'.)
3. Pogge's argument
    a. Reconstruction of the argument
        i. **(Unjust harm claim)** Several aspects of the global economic order *unjustly harm* the global poor.
        ii. **(Responsibility claim)** Governments of Western democracies are partially responsible this unjust harm.
        iii. **(Transmission principle)** If governments of Western democracies are partially responsible for unjust harms, then citizens of Western democracies are also partially responsible for the same unjust harms.
        Therefore,
        iv. **(Conclusion: *)** Citizens of Western democracies are partially responsible for unjust harm to the global poor.
    b. If (*) is true, then citizens of Western democracies presumably owe duties of justice towards the global poor, of two types:
        i. Reparation;
        ii. Duties to change the system (to stop the unjust harm).
    c. Elaboration of *Unjust Harm claim* and *Responsibility claim*
        i. The international resource and borrowing privileges
            1. Both of these concern the privileges that foreign nations assign (for specific purposes) to *whoever can gain de facto control of a country's territory*, without any constraints based on the political legitimacy of that control.
            2. The international resource privilege: Whoever is able to seize control of a country's natural resources is able to sell those resources to foreign (e.g. US, UK) customers, *and international law recognises and upholds the resulting property rights.*[12]
            3. The international borrowing privilege: Whoever has de facto control of a country is able to borrow money from international lenders (e.g. foreign governments), and international law upholds an obligation of the borrowing *country* (not merely the borrowing *person*) to pay back the resulting debt.
            4. How these privileges harm the global poor
                a. Grants effective ownership of resources to a country's political *leaders*, rather than to the country's *inhabitants as*

[12] Further (book-length but very accessible) discussion: Wenar, 'Blood Oil' (OUP 2016).

a whole. This impoverishes those inhabitants (relative to that alternative).

b. The political leaders in question are often 'authoritarians' who treat their own citizens terribly, and (in particular) in ways that drastically reduce the prospects for in-country economic development. In these cases, the effects of the international resource and borrowing privileges are
    i. To increase the incentives for such authoritarians to seize power.
    ii. To empower such authoritarians to stay in power without grass-roots political support, via sometimes brutal suppression of dissent.
    iii. Illustrative example: Wenar's correlation between authoritarian leaders' survival of the Arab Spring and *oil* income per capita. (*Blood Oil*, p.32)

ii. Protectionist economic subsidies and tariffs
    1. Rich countries often impose ("protectionist") import tariffs on goods imported from poor countries, and/or subsidise domestically produced goods, with the express purpose of decreasing the competitiveness of poor-country imports relative to domestic production.
    2. In the absence of these protectionist measures, there would be many more and better jobs in poor countries.

iii. Intellectual property rights protection
    1. The "Trade-Related Aspects of Intellectual Property Rights (TRIPS) Agreement": poor-country consumers of medicines must respect patents, and pay the same prices as rich-country consumers. This has the result that vast numbers of poor-country consumers are priced out of the market, even though they *would* be able to pay the marginal *production cost* of the medicines.[13]

iv. Installing and supporting malevolent dictators in poor countries
    1. E.g. the Shah of Iran, Saddam Hussein

v. Arms sales

vi. Western companies paying bribes to foreign officials
    1. Wenar (*Blood Oil*, pp.271-4) reports that this practice is now widely outlawed.

4. Objection: does this amount to *unjust harm*?
    a. Harm
        i. What the Pogge/Wenar discussion most directly supports is the claim that *the global poor would be a lot better off under various alternative* possible economic orders.
        ii. Comparative harm (definition): if P would be better off under some other state of affairs S' than under S, then a choice of S over S' (comparatively) *harms* P.
        iii. But not all comparative harm is unjust.

[13] Pogge, 'The health impact fund: Better pharmaceutical innovations at much lower prices'. Available online via https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1431180

1. Example: I give you £1000, rather than £1001.

iv. Natural suggestion: in any given decision situation C, there is some normative baseline state of affairs S* (which may or may not be available in C). Choosing S in C harms P iff P is worse off in S than in S*.
    1. Even if my only two options (for some reason) are to give you £1000 or £1001, the normative baseline is that I give you nothing.

v. How should the normative baseline be specified, for the purpose of evaluating 'the global economic order'? Some obvious suggestions:
    1. A particular ideal or minimally-acceptable state of affairs
    2. Some form of no-interaction state of affairs (e.g. no interactions between the affluent and poor countries, either (i) across the board or (ii) in the domain under discussion).

b. Injustice

i. Theories of outcome justice
    1. 'Theories of distributive justice' are normally theories of which distributions (of resources, or well-being, or something else) are just. These theories take justice/injustice to be properties of *outcomes* of the social/political/economic system.
        a. E.g. utilitarianism, prioritarianism, egalitarianism, sufficientarianism, …
    2. *If* one is willing to assume some such theory, it is easy to establish that the current global order is unjust – since its outcome is massively more unequal than necessary, denies far more people sufficient means to a minimally decent life, etc.
    3. However, Pogge (apparently?) wants to establish his conclusion without assuming an outcome-theoretic approach to justice.

ii. Procedural approaches to justice
    1. Basic idea: The primary objects of evaluation for justice/injustice are *procedures*.
        a. Outcomes can be labelled just/unjust only in the sense of having resulted from a just/unjust process.
    2. Usual form this takes: 'libertarianism'.
        a. Very roughly: there are various things one must not do to others (e.g. steal their legitimate property, assault them except in self-defence, take more than one's fair share of natural resources). But provided one acts within these constraints, one has fulfilled the demands of justice.
        b. Thus a libertarian holds that "human rights and justice [involve] solely negative [basic] duties" (Pogge, *World poverty and human rights* (WPHR), p.13).
        c. Libertarianism is often criticised for its lack of attention to outcomes (e.g., when it resists forced taxation to fund public goods).

iii. Dialectical situation
    1. Pogge *in fact believes* an egalitarian outcome-based theory of justice. But he does not want his argument to *rely* on this theory, because he wants his argument to appeal to the widest possible audience. (Often misunderstood.)

c. Injustice and harm
    i. The procedural approach to justice seems(?) to line up with a claim that comparative harm *relative to a no-interaction baseline* is unjust.
    ii. This makes Pogge's list of examples puzzling, though:
        1. Some of the aspects of 'the global order' that he cites do seem to amount to unjust harm on this account (the international resource and borrowing privileges, [installing malevolent dictators], arms sales).
        2. But others don't (intellectual property rights, protectionist economic measures).
    iii. Tan focusses on the second collection of examples, and concludes that Pogge's argument (for the consistency of his unjust-harm claim with the "libertarian tenet") fails.
d. Pogge's reply to Tan
    i. Pogge's argument does in fact rely on an outcome-based notion of justice, but a weaker one than his own "global egalitarianism": it assumes that "an institutional order is unjust if it foreseeably produces a substantial and reasonably avoidable human rights deficit", including deficits in "positive" human rights (notably the right to subsistence).
    ii. This condition for (in)justice supplies the baseline for evaluations of claims of unjust harm.
    iii. Justice requires us *not to impose* a global order that foreseeably commits unjust harm. (Note that this is a negative duty, insofar as 'imposing a global order' is action rather than inaction.)
    iv. This is the duty of justice that 'we' are guilty of violating.
e. Comments
    i. Pogge's reply undermines the point of the distinction between positive and negative duties. The same 'cheap trick' could be used to reclassify any archetypal positive duty as effectively a negative one, via institutions/'orders'. (E.g. "One violates no negative duty in *not giving to charity*, but one does in *imposing an order* according to which people (including ourselves) fail to give to charity.")
    ii. Pogge does not need to make this move anyway. If he appeals only to the international resource/borrowing privileges etc. (and not e.g. protectionist economic measures), he can argue more straightforwardly that Western governments commit relevant unjust harms.

5. The transmission principle
    a. Suppose that various actions of the *governments* of Western countries are indeed unjustly harming the global poor, along (at least some of) the lines Pogge sketches. To what extent, and how, would it follow that "we" owe duties of justice (including reparation) towards the global poor?
    b. Objection 1: I am not the government, and *I* didn't commit the unjust harms in question. Thus, my government owes duties of justice/reparation to the global poor (and thus ought to e.g. increase appropriate development assistance, and press change the global rules), but I as a private citizen do not.
    c. Reply: Your government was acting as your <u>agent</u>. You bear responsibility for what your agents do on your behalf.

        i. Pogge's example (WPHR, ch. 3 ("Loopholes in moralities")): You own an apartment building, currently occupied by long-term elderly tenants, and running a modest profit. You have the option to convert it into luxury flats. This would be more profitable, but would destroy the existing community (etc.) as the existing tenants could not then afford to remain. Realising that it would be unethical *directly* to oversee the conversion yourself, you hire a lawyer, whose brief is to maximise his interests (and who, therefore, foreseeably carries out the conversion).
    d. Objection 2: The government was not acting as my agent, because I did not support the actions in question, or any brief that would have entailed those actions.
    e. Reply: You benefit from the actions in question. You bear responsibility for wrongs from which you benefit.
    f. Objection 3: It is not in general true that one bears responsibility for wrongs from which one benefits. [14]
        i. Example: Data from Hiroshima inform studies of how much radiation is safe for humans. Thus, each of us benefits from the injustice perpetrated against Hiroshima, every time we go for an x-ray etc. But we do not *as a result of this* bear any responsibility for the Hiroshima bombings.
        ii. (Significant) concession: Benefitting from injustice may increase the stringency of one's duties *of beneficence* towards the victims of that injustice.
    g. Reply: you (not only benefit from, but also) contribute *to* the injustices in question. You bear (partial) responsibility for injustices to which you contribute.
        i. 3 ways one might relevantly 'contribute to' an injustice:
            1. 'Perpetuating injustice': preventing justice from being restored, once it has been committed
                a. E.g. keeping stolen goods that one has been involuntarily given, rather than returning them to one's owners
            2. 'Enabling injustice': encouraging others to initiate and maintain the injustice
                a. E.g. buying stolen goods
            3. 'Benefitting from injustice *at others' expense*'
                a. E.g. taking advantage of the victim's destitution to drive exploitative bargains
    h. Objection 4: None of these ways of 'contributing' applies in the cases Pogge discusses. (?)
    i. Last ditch: "Civic responsibility"
        i. Personal responsibility: Responsibility for the actions you yourself take
        ii. Civic responsibility: "We can legitimately hold people accountable to redress wrongdoing that they did not themselves commit by pointing to their responsibilities as members of a society that did commit wrongdoing."[15]
            1. This should not be a *sui generis* principle, though: if it is correct, what are the more fundamental principles from which it is derived?

[14] Anwander, "Contributing and benefitting: Two grounds for duties to the victims of injustice."
[15] Satz, 'What we owe the global poor'

Foundations of Effective Altruism
Frank Arntzenius and Hilary Greaves
Hilary Term 2017
Week 3: Making a difference

1. Background: Consequence-based reasons
    a. Consider a case in which the foreseeable result of *a <u>group</u> of people acting together in a certain way* is some good/bad outcome (e.g. that someone would be importantly benefitted/harmed).
        i. Q: What reasons for action do *individuals* have?
        ii. "Consequence-based reasons": reasons *deriving from the good/bad that is caused* by the group's action.
        iii. (If consequentialism is true, then all moral reasons are consequence-based.)
    b. "Share of the total view" (simplest case): If the group action consists of n people performing identical actions, then each individual has consequence-based reason, whose strength corresponds to 1/n of the total benefit/harm, for/against making her contribution.
    c. Against the "Share of the total" view
        i. Suppose that
            1. 10 people are trapped by rising waters
            2. There are 5 potential rescuers, including myself
            3. The cooperation of at least 4 rescuers suffices for the success of the rescue
            4. I know that the other 4 rescues will participate
            5. I can *either* participate in the group rescue, *or* act separately to save an additional, 11th person (but not both).
        ii. Then the "share of the total" view implies that I have stronger consequence-based reason to participate in the group rescue. But this is false.
    d. A more plausible view:
        i. The strength of my consequence-based reason for doing X rather than Y is the amount by which the consequences would be better if I did X than they would if I did Y (taking into account what others will in fact do).[16]
            1. Application to climate change: look at *marginal* rather than *average* damage
    e. Puzzle cases: sometimes it seems that a group has a consequence-based reason for acting in a certain way, while no member of the group has any consequence-based reason for playing their part. What to say about individuals' reasons, in such cases?
2. Some cases in which it might seem that no single individual makes a difference, although a sufficiently large group of people each performing identical actions does
    a. Some real-life cases

[16] This is *roughly* the view that Parfit labels "C6" (Reasons and Persons, section 25), although "C6" is phrased in terms of the amount by which a given act benefits/harms someone, rather than the strength of the agent's consequence-based moral reason for doing it.

i. Vegetarianism: 'my purchasing decisions make no difference to the number of animals killed'
ii. Climate change: 'individual emissions make no difference to climate damages'; 'individual travel decisions make no difference to emissions'
iii. Voting: 'no individual vote makes any difference to the outcome'
iv. Systemic change: 'no individual's role in the campaign makes any difference to the campaign's success/failure'

b. Some hypothetical cases
   i. The harmless torturers (Parfit, R&P, p.80): A victim is wired up to a device with settings from 0 to 1000. At setting 0 the victim feels no pain; at setting 1000 she experiences excruciating agony. Each of a thousand torturers presses a button, thereby turning the setting up by one. As a result, the victim suffers agony. But none of the torturers (individually) makes the victim's pain perceptibly worse.
      1. Similar case: drops of water (Glover)[17], [the self-torturer (Arntzenius and McCarthy)]
   ii. Vegetarianism (Kagan): The butcher orders another 25 chickens, and thus the farm kills another 25 chickens, every time the 25th chicken is sold. But it's very likely that any named individual's chicken-buying action has *no* effect on the number of chickens killed.
   iii. Firing Squad (Parfit): Two soldiers simultaneously shoot Joe. Either one of their bullets would have sufficed to kill.

3. The collectivist's reaction to these cases
   a. Individualism: All consequence-based reasons (including reasons applicable to groups) supervene on individuals' consequence-based reasons.
   b. Collectivism: the negation of individualism: Sometimes there are consequence-based reasons for a group to act in a certain way, even though there are no *consequence-based* reasons for the *individual members* of that group to play their respective parts in the group action in question.
      i. There may be other (non-consequence-based) reasons for individuals to act that way.
         1. Independent of the reasons that apply to the group: e.g. expressive action.
         2. Derivative from the reasons that apply to the group (? - controversial)
      ii. Or there may not (in which case we might hope that individuals are irrational).

[17] Drops of water case: Many men lie in the desert, suffering from extreme thirst. A large number of altruists have a pint of water each. If an additional pint is poured into a water-cart, each wounded man would get one extra drop of water. But "Even to a very thirsty man, each of these extra drops would be a very small benefit", and "might even be imperceptible".

c. Some critics who believe collectivism accuse the EA movement of having an (undesirable) 'individualist bias'.

4. Interlude: Objective and subjective goodness[18]; expected value
    a. The objectively best action is the action that would in fact lead to the best outcome.
    b. Objective ought: one ought to choose the objectively best action (when other things – considerations of promise-keeping, personal cost to the agent, etc. – are equal.)
        i. Problem: Usually we don't and can't know, at the point of decision, which action is objectively best. (So the objective ought is not action-guiding/seems inappropriate to blameworthiness.)
    c. Subjective ought (naïve version): one ought to choose the action that one *believes* to be objectively best.
        i. Problem: There won't always be any such action. (E.g. credences 40%, 30%, 30% that (respectively) action A, B, C is objectively the best.)
    d. Subjective ought (mark 2): one ought to choose the action X such that one's credence that X is objectively best is as high as possible.
        i. Problem: This gives (intuitively) the wrong answer in (e.g.) cases in which actions that might be best also might be very bad, while actions that certainly won't be best also certainly won't be too bad ("Jackson cases"/"mineshaft cases").
        ii. "3 pills": Sally has a mildly painful condition. Either pill A or pill B will cure the condition completely, but the doctor doesn't know which: there's a 50% chance that it's A, and a 50% chance that it's B. Whichever pill would not cure would kill. Pill C will (certainly) provide partial pain relief (only).
            1. Intuitive datum: The doctor (subjectively-)ought to prescribe pill C.
        iii. "Mineshaft": There are 100 miners trapped either in shaft A or shaft B (with 50% probability each way), and none in shaft C. If you divert the water into A or B, then any and all miners in that shaft will die. If you divert the waters into C, then ten of the miners will die, but 90 will be saved.
            1. Intuitive datum: you (subjectively-)ought to divert the water into C.
    e. Subjective ought ("correct" version): one subjectively-ought to maximise <u>*expected*</u> value, <u>i.e.</u> the probability-weighted sum of possible values (where the probabilities are either the agent's credences, or her evidential probabilities).
        i. (Worry: I don't always know my credences/evidential probabilities, so this too is sometimes not action-guiding.)

5. Taxonomy (details to be filled in in subsequent sections)
    a. Consider the graph <u>of total harm done</u> against <u>number of contributors</u> (or, more generally, size of total contribution). There are various types of case…
    b. Step function with unknown location (Kagan: "triggering cases")
        i. Structure: There is a small chance that the agent's action inflicts a *large* harm[19], and a large chance that it inflicts zero harm. The <u>*expected* harm</u>

[18] Some references for the literature on objective vs subjective oughts?

[19] Throughout, analogous points apply in cases that involve benefits rather than harms (e.g. drops of water). I discuss only the "harms" versions for simplicity of exposition.

inflicted is moderate (but large enough for the consequence-based reasons for action to be significant).

ii. Examples: Vegetarianism, taking a transatlantic flight, systemic change

iii. The agent has a consequence-based (subjective) reason not to inflict the harm.

c. Continuous or effectively-continuous function: (Kagan: "imperceptible harm cases")

i. Structure: Each agent inflicts an *extremely small*, but real, harm.

ii. Examples: drops of water, 'harmless' torturers, individual emissions

iii. The agent has a consequence-based (subjective and objective) reason not to inflict the harm, since the harms are (known and) real: any sense in which these harms are nonetheless "imperceptible" is normatively irrelevant.

d. Step function with known, 'past' location (overdetermination cases)

i. Structure: Harm is inflicted if at least n people make their contributions, but the agent knows that at least n *other* agents will contribute. Thus the individual's action genuinely and foreseeably makes no difference, *given* the actions that others (foreseeably) will take.

ii. Example: Firing squad

iii. The individual has no consequence-based reason to refrain from participating in the collectively-harmful action.

iv. Puzzle: How to reconcile this with

1. The fact that the collective does have such a reason?
2. The intuition that the individual does something wrong?

6. More on "triggering cases"

a. Suppose that

i. Every nth contribution causes ('triggers') a harm of size B.

ii. The agent has no information about how many other contributions have occurred.

b. By a 'principle of indifference'[20], the agent should have credence 1/n in each of the following propositions: The total number of contributions others will make {is an exact multiple of n, is (an exact multiple of n] plus 1, is (an exact multiple of n) plus 2, …, is (an exact multiple of n) plus (n-1)}.

c. In particular, the agent should have credence 1/n that her own act will trigger a harm of size B, and credence (n-1)/n that her own act will trigger no harm (or benefit).

d. Thus the *expected harm* done by her action is B/n.

i. In many cases (including Kagan's), this will be exactly equal to the agent's intuitive contribution (i.e. the approach defended here will coincide *in these cases* with the naïve 'share of the total' view).

7. More on "imperceptible harm cases"

a. Consider again the following claim: "In the case of the 'harmless' torturers, each individual torture inflicts a real harm on the victim."

[20] For a brief overview/discussion of this principle, and further references, see sections 3-4 of my paper 'Cluelessness'. (http://users.ox.ac.uk/~mert2255/papers/cluelessness.pdf)

b. Objection: these would-be harms are so small as to be imperceptible, and there cannot be imperceptible (hedonic) harms. Argument for this:

(P1) If two states feel the same to the patient, then those states are hedonically equally good/bad.

(P2) If the difference between two states is imperceptible, then those states feel the same to the patient.

(P3) The difference between two adjacent states in the 'harmless torturers' case is imperceptible.

Therefore,

(C1) Adjacent states are equally hedonically good (or bad). (From P1, P2, P3)

(P4) If adjacent states are equally hedonically good/bad, and no non-hedonic considerations are relevant, then no individual torturer harms the patient.

(P5) No non-hedonic considerations are relevant.

(C2) No individual torturer harms the patient. (From C1, P4, P5)

c. Reply: The argument trades on an equivocation. Depending on how one makes 'imperceptible' precise, either (P2) or (P3) is false.
    i. Thought-experiment (Arntzenius and McCarthy, section 2):
        1. Suppose the patient experiences a long sequence of randomly-ordered dial settings from the 0-1000 spectrum, and attempts, as each dial setting is experienced, to describe how painful the corresponding state is. Tabulate the responses.
        2. For each dial setting, there will be a certain frequency profile of responses, and (since there will be a difference in the frequency profiles corresponding to states 0 and 1000) there *must* be at least one pair of adjacent dial settings corresponding to different frequency profiles of responses.
        3. But the patient is issuing descriptions only on the basis of how the states *feel* to him. Therefore, for at least one adjacent pair, there must be a difference in how the states in that pair feel (and this remains true even if the procedure described in this thought-experiment is merely counterfactual).
d. (Terminological) question: Are the differences in question "perceptible"?
    i. First sense of "perceptible": A difference is perceptible iff the probability profile of a subject's descriptions of how that state feels are different.
        1. In this sense, the difference between at least one pair of adjacent states in the 'harmless torturers' case is perceptible. (P3 is false)
    ii. Second sense of "perceptible": A difference between two pain states is perceptible only if, If subjected to the two pain states in question consecutively and asked "can you tell which of these is more painful", the patient sometimes replies "yes".
        1. In this sense, the differences between adjacent is (perhaps) imperceptible; but if so, we are forced to conclude that states that are only imperceptibly different can *feel* different. (P2 is false)

8. Overdetermination cases

a. Consider again 'Firing squad'. Do the individuals do anything wrong?
    i. This depends on *why* the individuals shot (and/or: under what other circumstances they would/would not have shot), which has not yet been specified.
b. Case 1: Each soldier would shoot regardless of what he believed about the other's action, because he is keen to be part of the execution.
    i. The soldiers' *motivations* are criticisable in consequence-based terms (modally)[21], although their *acts* are not.
c. Case 2: The first soldier shoots because he does not care (enough) whether or not the victim dies, and wants to avoid the minor punishment that he would suffer for disobeying orders. The second soldier shoots because he correctly believes this about the first soldier, hence knows that his additional bullet will not make a difference, and wants to avoid the minor punishment; but the second soldier wouldn't have shot if he'd thought the first might well not shoot.
    i. The first soldier's motivations are criticisable in consequence-based terms, as above. The second soldier is not morally criticisable in consequence-based terms (for his act or for his motivation).
d. Case 3: The two soldiers previously discussed what to do in this situation, and agreed that both would shoot. Each soldier is *now* such that he wouldn't shoot if he thought the other party would not, but, thanks to the agreement, knows that in fact the other will shoot. Thus both shoot (to avoid minor punishment).
    i. The soldiers had consequence-based reason, at the earlier time, not to make the agreement to shoot. *Making this agreement* was wrong in straightforwardly consequence-based terms.
e. Case 4: There was no earlier agreement, but each soldier correctly *guesses* that the other will shoot. Thus both shoot (to avoid minor punishment).
    i. This is a coordination problem:

	A	B
A	Best	Worst
B	Worst	Second-best

    ii. Another example with this structure: meeting for lunch at the curry house vs at the pizza place
    iii. Note that opting for the action that leads to the *second*-best outcome is peculiar (irrational?) behaviour in problems with *this* structure. (Although choosing the second-best can be rational in slightly different cases – what if there were 2 curry houses, but only one pizza joint? The idea of a 'Schelling point'.)
    iv. If the case really is as described, then
        1. The soldiers are immune to *moral* criticism

[21] Cf Pinkert's "modally robust act consequentialism" (Pinkert, 'What if I cannot make a difference (and know it)')

2. They may be liable to *rational* criticism (for failing to choose the Schelling point).
3. The outcome is a result of unfortunate coordination failure.
4. Can we blame the *pair* of soldiers for the victim's death?
    a. *If* reasons apply to the pair at all, then the pair (i) does have strong consequence-based reason not to shoot, and (ii) did not act on that reason, so (iii) can be blamed.
    b. But given the failure of coordination, the pair of soldiers is arguably not an agent (did not act at all), and not an appropriate subject for criticism/blame/etc., or a subject of reasons.[22]

9. Application: Climate change
    a. Sinnott-Armstrong claims that each individual's emissions foreseeably make no difference to climate damages, so that there is no reason *based on harm caused* for e.g. refraining from recreational driving.
    b. His reasons for this claim amount to a catalogue of mistakes, though:
        i. First mistake: False dichotomy
            1. "Global warming will occur even if I do not drive just for fun"; "my individual act is neither necessary nor sufficient for global warming".
            2. Reply: 'Global warming', and climate damages, are matters of degree.
        ii. Second mistake: Neglecting 'imperceptible' harms
            1. "Greenhouse gases… are perfectly fine in small quantities…. The problem emerges only when there is too much of them. But my joyride by itself does not cause the massive quantities that are harmful."
            2. Reply: Disambiguate. Greenhouse gases "are perfectly fine in small quantities" in the sense that there would be no climate damage if total emissions were low. But they are not "perfectly fine in small quantities", *even at the margin*, when existing total emissions are already high.
        iii. Third mistake: neglecting triggering effects
            1. "You might think that my driving… raises the temperature of the globe by an infinitesimal amount. [B]ut even if it does, my exhaust… does not cause any climate change at all. No storms or floods or droughts or heat waves can be traced to my individual act of driving."
            2. Reply: This is just like Kagan's chickens.
        iv. Fourth mistake: Empirical confusion
            1. Sinnott-Armstrong seems to think that climate change is an overdetermination case. (?)
                a. But there is just no reason for thinking this, as an empirical matter. (The fact that the climate system is extremely complex tells *against* the overdetermination claim.)

[22] Cf. he discussion of 'collectives' in Collins, "Collectives' duties and collectivisation duties", sections 1 and 2.

Foundations of Effective Altruism
Frank Arntzenius and Hilary Greaves
Hilary Term 2017
Week 4: Career choice

1. Introduction: Earning to give (EtG)
    a. Paradigm ethical careers, on the common-sense view: charity work, social work, state-school teaching, medicine…
    b. MacAskill's suggestion: take a high-paying job (in e.g. finance) and donate most of your earnings to the most cost-effective charities. ("Earning to give")
        i. Claim: EtG is usually *ethically preferable* to the paradigm ethical careers, even if the high-paying job in question is "morally controversial".
        ii. Further, this is so even by the lights of those who argue for the paradigm ethical careers (via "making a difference").
    c. Considerations in favour of EtG:
        i. Financial discrepancy argument: Because of the salary difference, by taking the lucrative job you can make it the case that more charity workers are hired.
        ii. Fungibility argument: you can easily *direct your money to* the most cost-effective charities (insofar as you know which those are), but it's much harder to *get a job working for* the most cost-effective charities.
        iii. Uncertainty argument: you can *switch your donations* to more cost-effective charities as new information (about cost-effectiveness) comes in, but it's much harder to *switch jobs* to a more cost-effective charity.
        iv. Replaceability argument:
            1. If you don't take the charity job, someone else *more-or-less like you* will take it instead. If you don't take the EtG route, someone else *who will not donate [well]* will take the lucrative job instead.
            2. Thus the *outcome* of your taking the charity job (and a non-philanthropist taking the lucrative job) is worse than the *outcome* of your taking the lucrative job (and a person basically like you taking the charity job).
                a. Setting aside considerations of what the person you *do* displace goes on to do instead…
                b. This argument seems to assume that in neither case do you make any difference to how many workers in each industry are hired. That assumption is false.[23] But a similar argument probably goes through.
    d. Two cases to consider: EtG via a "morally innocuous"/"morally controversial" career
        i. Taxonomy: Say that a career is
            1. *morally innocuous* if there are *no* strong non-consequentialist reasons against doing the work in question.

[23] O'Keeffe-O'Donovan, "What does economics tell us about replaceability", https://80000hours.org/2014/07/what-does-economics-tell-us-about-replaceability/

        a. E.g. journalism, blue-skies research, shop manager
    2. *morally controversial* if there *are strong* non-consequentialist reasons against doing the work
        a. E.g. (arguably), petrochemicals, arms, speculative finance
    3. *morally reprehensible* if there *are extremely strong or overwhelming* non-consequentialist reasons against doing the work
        a. E.g. hit man, child trafficker, concentration camp guard
  ii. MacAskill's "weak claim" (resp. "strong claim"): EtG via a morally innocuous (resp. morally controversial) career is typically ethically preferable to pursuing a paradigm ethical career.
  iii. NB MacAskill "[does] not wish to defend philanthropy through morally reprehensible careers" (p.274).
2. Objection: Two senses of 'making a difference'
    a. Technical sense: The difference I would make by doing X rather than Y (e.g., EtG rather than working for a charity) is the difference between *the state of affairs that would result if I did X* and *the state of affairs that would result if I did Y* (given other people's actual dispositions, and other aspects of the way the world is).
    b. Ordinary-language, causal sense: The difference I would make by doing X consists in the effects that *I myself ('directly') cause* if I do X.
        i. Thus, e.g., if I work as a bed-net distributor, I make the difference to the recipients I give bed-nets to (despite the fact that if I had not made that difference, someone else would have made the very same difference to the very same people).
        ii. Hard/messy to make this precise – but clear enough in practice?
    c. MacAskill argues for the claim that EtG makes more difference in the technical sense. But paradigm ethical careers arguably 'make more of a difference' in the ordinary-language sense than EtG *careers*.
        i. Although EtG *donations* are a different story…
    d. Thus (pace MacAskill) the defender of paradigm ethical careers need not be making a mistake *by her own lights*.
    e. It's a further question whether her lights are the right ones, though.
        i. Is it 'morally self-indulgent' to be concerned with the difference you make in the ordinary language sense, rather than in the technical sense?
3. Reconstruction of MacAskill's argument for the weak claim

    (P1) EtG typically makes much more of a positive difference, in the technical sense, than following a paradigm ethical career. (Premise; illustrated by the arguments from financial discrepancy, fungibility, uncertainty and replaceability)

    (P2) If one action makes much more of a positive difference than another in the technical sense, and there are no strong non-consequentialist reasons against the first, then the first is ethically preferable to the second. (Premise; NB this ignores the possibility of non-consequentialist reasons *in favour of* paradigmatic 'ethical careers')

    (P3) There are no strong non-consequentialist reasons against EtG via a morally innocuous career. (By definition)

    Therefore,

(C) EtG via a morally innocuous career is typically ethically preferable to following a paradigm ethical career. (From P1, P2, P3)

4. Interlude (I): Jim and the Indians[24]
    a. Q: What should Jim do?
    b. Utilitarianism: (1) The correct answer is 'Kill one Indian'. (2) This is obvious, because nothing except the number of resulting deaths is morally relevant.
    c. Absolutist deontology: Don't shoot, because there is an absolute side-constraint against killing.
    d. Williams (roughly): (1) is correct but (2) is not (therefore utilitarianism is false).
5. Interlude (II): Integrity, and George the chemist[25]
    a. Q: What should George do?
    b. Utilitarianism: George should accept the job.
    c. Williams: Probably, George should not accept the job.
        i. Since pacifism is one of George's deep commitments, George cannot accept the job without violating his own integrity, and this is a weighty moral consideration.
        ii. Utilitarianism is unattractive because, in requiring agents to assign no more weight to their own projects/commitments than to those of others, it "alienates" an agent "from his actions and the source of his action in his own convictions". (This is "the integrity objection to utilitarianism.")
6. Interlude (III): Consequentialism, doing/allowing and the 'doctrine of double effect'

---

[24] "Jim finds himself in the central 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 protestors 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 of the soldiers to threat, but it is quite clear from the set-up that nothing of the sort 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?"

[25] "George, who has just taken his Ph.D. 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 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, come to that, 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 to get 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?"

a. Consequentialism is notoriously indifferent to the *more detailed nature of the relationship* between an action and the resulting state of affairs.
b. E.g. 'Organ harvesting': A doctor has six patients, one with a minor complaint, and five urgently needing organ transplants. The doctor can either allow the five to die from organ failure, or can kill the first patient in order to transplant his organs into the five.
c. Consequentialism either says that the doctor should kill the one, or (if not) says that that the only reasons why not arise from contingent, instrumental effects (the story would get out and would undermine trust in the medical system, etc.) This strikes many people as wrong.
d. First distinction: Action/inaction
    i. The first patient would die as a result of some positive *action* of the doctor's. The five would die as a result of *inaction*. Suggestion: Moral requirements against *action* are typically more stringent than those against *inaction*.
    ii. Objection: Some inactions are morally on a par with actions: Foot's actor
e. Second distinction: Doing/allowing[26]
    i. The choice is between *killing* one vs. merely *allowing* five to die. Suggestion: Duties not to *do* harm are typically more stringent than duties not to *allow* harm.
    ii. Why this isn't the same as action/inaction[27]
        1. Analysis: We have a notion of some 'default' sequence of events being "somehow already in train". *Doing* is intervening in that sequence of events, while *allowing* is leaving it to run its course.
        2. Explanation of the difference between action/inaction and doing/allowing: The 'default' sequence of events might include some actions on the part of the agent in question.
f. Third distinction: Intending/merely-foreseeing (and DDE)[28]
    i. If the doctor allows the five to die, she does not *intend* their deaths, although she *foresees* them. But if she kills the one to save the five, she *intends* the death of the one.
    ii. Suggestion ("the doctrine of double effect (DDE)"): one usually must not *intentionally* harm others, but one often may *foreseeably cause* harm to others, provided that the harm is *merely* foreseen (i.e. not also intended).
    iii. Another example: Terror bombing vs. tactical bombing
        1. Terror bombing: In order to end the war early, generals decide to bomb population centres, hoping that the resulting civilian deaths demoralise the enemy.
        2. Tactical bombing: In order to end the war early, generals decide to bomb the enemy's munitions factories, although they know that civilian deaths will be among the results.
        3. Suggestions:

[26] See also https://plato.stanford.edu/entries/doing-allowing/
[27] Foot; see also Kagan, *The limits of morality*, ch. 3, esp. p.94; and cf. week 2
[28] See also https://plato.stanford.edu/entries/double-effect/

            a. The civilian deaths are intended in 'terror bombing', but are merely foreseen in 'tactical bombing'.
            b. Tactical bombing is permissible, but terror bombing is not. (Or anyway tactical bombing is *better*.)
        iv. More (similar) examples: Trolley problems (the switch vs the fat man), euthanasia vs. palliative care
        v. Analysis (roughly – but see e.g. Kamm)
            1. One of the foreseen effects is your *goal* (e.g. ending the war early). You *intend* an effect if it is a means to your goal. You *merely foresee* it if it is a side-effect of some means to your goal, or a further effect of the goal itself.
        vi. Objections to DDE
            1. Doesn't match common intuitions in all cases – cf. 'loop' version of the trolley problem[29]
            2. We can always redescribe the situation so that the harm counts as merely foreseen. ('I *intended* only that the fat man stop the train, not that he die…') Thus the verdicts of DDE are indeterminate.
    g. Objections to all 3 suggestions
        i. Whether one classifies a given case as action or inaction, doing vs allowing, or intending vs merely-foreseeing depends on a prior judgment about the permissibility of the action in question ("the Knobe effect"[30]). Thus the theories in question, even if correct, are uninformative.
        ii. No reason why any of these distinctions should carry moral weight. (Kagan, 'The limits of morality', esp. chapters 3 and 4)
    h. Foot's position: The doing/allowing distinction is a better way to deal with cases like 'organ harvesting' than DDE, because DDE has ridiculous consequences in the abortion case.
        i. But doing/allowing cannot recover the desired(?) verdict on terror vs. tactical bombing, euthanasia, etc.
        ii. Alternative reaction: DDE does not in fact justify the conclusions that Foot objects to. (The death of the foetus is not intended in either the hysterectomy or the abortion case that Foot discusses. Foot says that this move "would make nonsense of [DDE] from the beginning"- ?)
7. Returning to career choice… Argument for MacAskill's "strong claim"
    a. To defend the strong claim, one has to either
        i. Argue that the non-consequentialist reasons against the morally controversial career are *outweighed* by the consequence-based reasons in favour.
            1. Analogy: breaking a promise in order to save a life

[29] Thomson, "The trolley problem"; Kamm, "Intricate ethics", chapter 4

[30] Knobe, "Intentional action and side effects in ordinary language" and "The Concept of Intentional Action: a Case Study in the Uses of Folk Psychology"; Kagan, *The limits of morality*, pp. 101-5 (on the same phenomenon for doing/allowing).

ii. Argue that non-consequentialist reasons that would sometimes apply to similar career choices *don't apply* to the particular case under consideration. (MacAskill's strategy)

b. Harm-based reasons
    i. As in the case of 'make a difference', we can distinguish between an ordinary and a technical sense of 'harm':
        1. Technical sense: a choice of X over Y *harms* S iff S is worse off under X than under Y.
        2. Ordinary sense: need not line up with the technical sense. E.g., a gold bar dropping from the sky harms you if it lands on and breaks your foot, even if you end up better off overall (since you get to keep the gold bar).
    ii. A morally controversial career involves, inter alia, *harming* people (in the ordinary sense). There are strong non-consequentialist based reasons not to harm people (in this sense). Couldn't these outweigh the consequence-based reasons to EtG?
        1. Not if the EtG path leaves all the would-be 'victims' better off – e.g., you harm people as part of your career, but you harm them *less* than your replacement would have harmed the very same people.
            a. Analogy: 'Jim and the Indians' (Williams)
        2. And not if the harm is only *foreseen*, not *intended* – e.g., you would earn the same high salary in a way that didn't harm anyone if you could.
            a. Analogy: 'Tactical bombing', and other 'double effect' cases
            b. Real-life example: Schindler

c. Integrity-based reasons
    i. Analogy: 'George the chemist' (Williams)
    ii. MacAskill's reply
        1. The commitment e.g. not to work for petrochemicals, if that includes working for petrochemicals *in order to give*, is simply based on a *mistake*.
        2. Integrity does not recommend acting on a given commitment if that commitment is itself based on a mistake.
    iii. Real-life example: Engels

8. General worry with this argument-strategy
    a. All the same arguments apply equally (?) to "morally reprehensible" careers. But that would be a *reductio* (as MacAskill agrees).
9. A suggestion
    a. What the typical 'altruistic graduate' actually wants is to be engaged in work in which *her direct, day-to-day aims* are altruistic.
    b. What the 'EA critique' highlights is that
        i. this is not pure altruism, and perhaps
        ii. pure altruism is morally superior.
    c. But the extent to which non-consequentialist considerations should constrain the pursuit of pure altruism remain obscure.

Foundations of Effective Altruism
Frank Arntzenius and Hilary Greaves
Hilary Term 2017
Week 6: Conditional obligations

1. Structure of a "conditional obligation" case: For some acts A, B, C,
    a. A is morally optional (i.e., both A and not-A are permissible).
    b. Doing A gives rise to a further choice between B and C.
    c. The ranking of actions *in terms of [expected] goodness of their outcomes* is: A&B (best), A&C (middle), not-A (worst).
    d. B is required if A is done (i.e., A&C is forbidden), but
    e. A&C *would* be permitted *if* A&B were unavailable.
2. Examples (alleged)
    a. Kagan's parrot[31]: Optional to enter the burning building, but if one does, one must save the child rather than the parrot.
    b. Rulli's medical trial: optional to carry out the study overseas (as opposed to in the US), but if it is carried out overseas, it must involve active care rather than placebo in the control group.
    c. Kamm's disturbed visitor: optional to stay in town, but if one stays, one must offer full (not merely partial) emotional support.
    d. Sweatshops: Optional to employ poor workers, but if one employs them, one must offer them decent working conditions, pay etc.
    e. Kavka's slave child: optional to have a child, but if one has a child, one must do more than merely ensure the child's life is worth living.
    f. EA: Optional to give (at least beyond a certain threshold), but if one gives, one must give cost-effectively.
3. The existence of conditional obligations is inconsistent with:
    a. Any form of consequentialism (unless we gerrymander the theory of the good[32]) – whether maximising, scalar, or satisficing.
    b. The "moral free space view": if A is optional and if A&C would be permitted if only {A, A&C} were available, then A&C remains permitted when {A, A&C, A&B} are all available.
4. What's supposed to be puzzling about conditional obligations
    a. "Discrepant rankings objection": on a conditional-obligations view, the ranking of acts by goodness of outcomes is different from the ranking by normative status. How come? (Not really an *objection* – rather, a demand for explanation.)
        i. Rulli's key claim: a proper (Scheffler-style) understanding of how 'moral options' arise will also explain conditional obligations. (TBC.)
    b. "Easy exemption objection": it's easy to avoid the onerous obligations, by performing the worse act. Isn't something wrong with that?
        i. Reply: It's impermissible to perform not-A *motivated by a desire to avoid the onerous obligation to perform B*, although it's permissible to perform not-A

[31] Kagan, *The limits of morality*, p.16.
[32] For the project of gerrymandering the theory of the good in an attempt to get as many theories as possible to count as formally consequentialist, see e.g. Portmore, 'Consequentialising', and references therein.

motivated by desire to avoid costs directly associated with A. (This removes the 'perverse incentives' aspect of the problem.)

c. "Mutual agreement objection": in some cases, the parties could would mutually consent to the choice of A-and-C over not-A, on grounds of Pareto-superiority.
    i. Rulli's reply: In the cases in question, it would be impermissible for the agent to *offer* A-and-C, so the consent would not generate permissibility.

5. Interlude: Scheffler's theory[33]
    a. Motivations
        i. Maximising consequentialism (recall): one is required to produce the best possible state of affairs.
        ii. What this means in practice is fairly flexible, since it says nothing about the 'theory of the good': one could e.g. assign much higher weight to killings than to lettings-die, in one's ranking of outcomes. ('Consequentialising')
        iii. But the flexibility has limits. In particular, as long as the theory of the good is agent-neutral, maximising consequentialism denies that there is any agent-relativity. In particular, there can be no:
            1. Agent-centred restrictions: One is not always even *permitted* to maximise the good, since sometimes doing so would involve violating a 'side-constraint'. E.g. killing/lying/etc is forbidden (up to a point) even if it would lead to fewer killings/lies/etc overall.
            2. Agent-centred prerogatives: One is not always *required* to maximise the good, since one is *permitted* to give some priority to one's own family/friends/projects/etc. (e.g. George the chemist, again…)
        iv. Scheffler thinks that agent-centred restrictions are borderline-paradoxical, but wants to defend agent-centred prerogatives.
    b. Outline of the theory (suggested)
        i. Let X be the set of states of affairs.
        ii. Let V_{imp}:X→R be an *impartial* value function (so V(x)>V(y) whenever x is better than y *from a completely impartial point of view*) – e.g., the utilitarian one.
        iii. Let V_S:X→R be a *non-impartial* value function for agent S, tracking only (i) the well-being of persons that S has a special interest in (including S herself) and (ii) the pursuit/success of S's projects. ("The personal point of view")
        iv. Then S is required to [[maximise [the expectation value of] the linear combination $\lambda V_{imp} + (1-\lambda)V_S$, for some permissible value λ ($0 < \lambda_{min} <= \lambda <=1$).]]
            1. 'Permissible': Not permitted to assign *zero* weight to impartial value; permitted, but not required, to assign zero weight to S-value.

6. Applying "Scheffler" to conditional-obligation cases
    a. "Scheffler" on Kagan's parrot
        i. *Suppose* that
            1. Both the parrot and the child are complete strangers to the agent.

[33] Scheffler, *The rejection of consequentialism*.

2. None of the agent's "projects" favours saving the parrot over the child (or, for simplicity, vice versa).

ii. Then:

1. V_S assigns equal value to saving the child vs saving the parrot.
2. But V_{imp} assigns higher value to saving the child (we take it).
3. Therefore any permissible linear combination $\lambda V_{imp} + (1-\lambda)V_S$, with $\lambda>0$, will assign higher value to saving the child than to saving the parrot.
4. Therefore, according to "Scheffler's" theory, the agent is not permitted to save the parrot when saving the child is also an option.

iii. Suppose further that

1. The risk to the agent, in entering the burning building, is significant.

iv. Then:

1. V_S assigns lower value to *both* saving the child *and* saving the parrot than it does to staying outside.
2. It can easily happen that some permissible linear combination $\lambda V_{imp} + (1-\lambda)V_S$, with $\lambda<1$, assigns higher value to staying outside than to either saving the child or saving the parrot.

b. Rulli's (equivalent?) analysis: Consider pairwise comparisons between actions, one of which is *impartially* inferior to the other.

i. C = "the [impartial] cost of promoting the inferior outcome preferred by the agent rather than the optimal outcome".
ii. S = "amount of sacrifice to the agent in promoting the optimal outcome".
iii. N = "the additional weight moral options allow us to apply to the agent's interests in the matter".
iv. Options: iff C <= NS, then there is an option to perform the inferior act. It follows that there is an option to perform the impartially inferior act from the pair {save child, stay outside}, but not from the pair {save child, save parrot}.

7. Accounting for the attraction of the suboptimal action

a. Problem: The above analysis makes nonsense of the agent's *desire* to perform the suboptimal action rather than the impartially optimal one.
b. In more realistic cases, there is at least *something* at stake for the agent in the choice between the permitted act A&B and the allegedly-forbidden act A&C, in *some* sense of 'at stake'…

i. Medical trial: The placebo-controlled experiment would be more informative, or less expensive
ii. Disturbed visitor: Partial support would be less time-consuming than full support
iii. Slave child: The point was to create a slave…
iv. Parrot case: I'm a bird lover??

c. Scheffler-style analysis, acknowledging this attraction

i. Suppose the amount of sacrifice to the agent, in doing A&B rather than A&C, is (positive but) *modest*, while the sacrifice to the agent in doing A&C rather than not-A is *significant*.

    ii. Meanwhile, suppose that the three actions are *equally-spaced* in terms of impartial goodness.
    iii. For suitable ways of making this precise (see Appendix), no convex linear combination of agent's value and impartial value will rank A&C above both not-A and A&B: *either* the relative weight assigned to impartial value is sufficiently high that the convex combination prefers A&C to A&B, *or* the relative weight assigned to agent's value is sufficiently high that the convex combination prefers not-A to A&B; thus A&B is never permitted.
8. Accounting for the agent's preference for the suboptimal action *over just staying out of it*
    a. Problem: The above analysis only works if not-A has the highest S value. Thus it makes nonsense of the agent's desire to perform the suboptimal action *rather than the impartially even worse, but intuitively permitted one (not-A)*.
    b. To make genuine progress, we need to ask why the agent would want to do that (morality aside). Plausible guesses:
        i. Kagan's parrot: Either the parrot is the agent's pet, or the agent is a particularly avid bird-lover.
        ii. Rulli's medical trial: Carrying out the study overseas is cheaper than doing it in the US, and the placebo-controlled trial is more effective than the active-control trial.
        iii. Kamm's disturbed visitor: The agent has things she needs to do at home. It isn't convenient for the agent to go away for the weekend, nor to spend the *whole* weekend supporting the visitor.
        iv. Kavka's slave child: The agent badly needs the money that she could get by selling a child into slavery, but doesn't otherwise want a child.
    c. In all these cases, there is at least one plausible account of "the agent's well-being together with the pursuit and/or success of her projects" according to which, by the lights of that value scale, the forbidden option scores highest. (That's why it's the agent's own preference…)
        i. *If* V_S tracks that account, then the Scheffler-style theory is in principle incapable of recovering the claim that not-A is permitted while A&C is forbidden. (Both the impartial value function and the S-value function assign higher value to A&C than to not-A, so clearly any linear combination of them will do the same – thus not-A comes out forbidden (and A&C may or may not).)
        ii. Alternatives…?
    d. Tentative conclusion: Contra Rulli, a Scheffler-style theory *cannot* account for conditional obligations, in cases in which there is *something at stake* (broadly construed) for the agent to rationalise the agent's preference for A&C over A&B.
9. Other things that might drive (and justify?) the 'conditional obligation' intuition
    a. Kagan's parrot (and EA?): Not all elements of the 'personal point of view' are morally on a par. Morality cares more about well-being costs to the agent than about project costs. E.g. the agent's motives not to enter the building carry more moral weight than her motives to save the parrot once inside, relative to the weights the agent herself assigns to those motives.
    b. Medical trial, sweatshops: performing the action A puts you in a relationship that you wouldn't otherwise have had; what's morally objectionable is the conjunction of *having that relationship and* not treating the other party well.

i. Analogy (just about): It's optional whether or not to make a promise, but <u>if you make the promise you have an obligation to keep it</u>. Whether the state of affairs in which you make but break the promise is better than the state of affairs in which you don't make the promise in the first place is beside the point. (Cf. Frick, 'Conditional reasons and the procreation asymmetry')
ii. Rulli says that her conditional-obligation cases are *not* like promissory cases, in that in her cases, "no … contract/promise is constitutively involved in doing A." But perhaps this difference (for a narrow sense of contract/promise) is irrelevant?

c. Slave child: Similar to the medical trial (etc), except that the language of 'relationship' might be strained here?

10. Application to EA: Two cases to consider
    a. Gratuitous case: You have £1000 to give away. You don't think about cost-effectiveness, and just give it to a charity you like the sound of/the first charity you come across in your chosen cause area. This charity turns out to be 1000x less cost-effective than another charity working in the same cause area (e.g. HIV/AIDS treatment).
    b. Motivated case: You decide to give £1000 to a cancer research charity, because you have recently lost a relative to cancer, and you specifically want to support this cause. You donate to the most cost-effective cancer charity you can find, but you're aware that by the lights of any reasonable cause-neutral metric, some non-cancer charities (e.g. AMF) are 1000x more cost-effective.
11. Pummer on 'gratuitous worseness'
    a. Pummer's claim: You have a conditional obligation to donate to the most cost-effective charities *when failing to do so would be gratuitous*. (For the most part, Pummer does not discuss the motivated case, except to concede that his argument might "very plausibly" not apply to that case – but cf. section 10 of Pummer's article.)
    b. Analogy: the 'arm donor' case
    c. Pummer's general principle:
        i. "Avoid gratuitous worseness (weak): It is wrong to perform an act that is much worse than another, if it is *no costlier* to you to perform the better act, and if all other things are equal."
    d. Objections and replies
        i. Objection: If it's not wrong to do nothing, then there is no-one who is wronged when you do nothing, and therefore no-one to whom you owe an obligation to help. But then it follows that you violate no helping-obligation that you owe to anyone if you help less than you could, either. Therefore helping less than you could is not wrong.
            1. Reply (not Pummer's): Not all impermissible acts are wrong *in virtue of* wronging some particular person.
        ii. Objection: Suppose there were also an available act (A&D) that had the same cost to the agent as A&B/A&C, but whose benefit accrued to the agent, rather than to others. On the conditional-obligation account, A&D is permissible. But since "it's morally better to be kind than selfish", if A&D is permissible then A&C must also be permissible.

1. Reply: A choice of A&C may lead to a more favourable assessment of the *agent* than a choice of A&D, but this doesn't entail that we must make a more favourable assessment of the *act*. (?)

iii. Objection: If you can specially favour yourself, why can't you specially favour others too?
   1. Singer's objection: "If I am permitted to keep my $1000 rather than give it to the Against Malaria Foundation, presumably that means that I am permitted to, say, spend it on a cute haircut for Fido, my own dog, because I enjoy seeing Fido with a cure haircut. So why aren't I permitted to donate it to Cute Haircuts for Princeton Dogs, because I enjoy seeing dogs around my neighbourhood with cute haircuts? Or donate it to Cute Haircuts for Paraguayan Dogs, because I like the thought of dogs in that country having cute haircuts, even though I will never see them?"
   2. Reply: the relevant considerations here are quantitative. "As we move further and further away from the things you really care about… it becomes progressively less plausible that you are permitted to use your money in the specified ways. The basis of the moral option to do less good gradually disappears."
      a. But note that this reply gives up a significant part of the original "optionality about whether to give" claim.

e. Extending the argument to the motivated case
   i. Once Pummer's claim for the 'gratuitous' case is granted, it is straightforward to generalise to cases in which there is a *slight/moderate/etc* cost to the agent in doing *much* more good, and analogues of many of the same arguments will still apply.
   ii. Possible application to the EA case: "Considerations of cost to you [in the broad sense] may permit you to give some portion but not all of your "donation money" to charities closer to your heart, and perhaps then it would be wrong not to give the remainder to those that do much more good per dollar donated."

Appendix: How to outlaw the middle ground

Let x, y, z be actions. (Intended application, in Rulli's notation: x=A&B, y=A&C, z=not-A.) Suppose that the impartial ranking is x>y>z, while the agent's ranking (i.e. the ranking according to V_S) is z>y>x. Define D-values (difference values) as follows:
$D_1=V_{imp}(x) - V_{imp}(y)$; $D_2=V_{imp}(y) - V_{imp}(z)$; $D_3=V_S(z) - V_S(y)$; $D_4=V_S(y) - V_S(x)$
Then it is straightforward to show that provided $D_3/(D_2+D_3) < D_4/(D_1+D_4)$, no convex linear combination of V_{imp} and V_S ranks y above both z and x. (Either λ is sufficiently low – low weight to impartial value – that z is ranked above y, or λ is sufficiently high that x is ranked above y.)

Foundations of Effective Altruism
Frank Arntzenius and Hilary Greaves
Hilary Term 2017
Week 7: Population ethics and existential risk

1. Introduction: Extinction risk
    a. What we can reasonably expect if "nothing goes wrong"
        i. Some context:
            1. *Homo Sapiens* has already been around for 200,000 years
            2. The average mammalian species lasts for 1-2 million years
            3. The average historical frequency of mass extinction events is 1 per 100 million years
            4. The heating-up of the Sun will dry out the Earth in something over 1 billion years' time.
        ii. Thus: reasonable to expect sentient life on Earth to continue for at least (say) 1 million years.
        iii. There are *live possibilities* of continuation vastly beyond this (via colonisation of other planets). (Important insofar as the relevant parameter is *expected* survival time of sentient life.)
    b. Some sources of extinction risk
        i. Nuclear war
        ii. Asteroid impact
        iii. Synthetic biology
        iv. Artificial general intelligence
    c. Some mitigation options
        i. Source-specific: nuclear disarmament, asteroid deflection, AI safety
        ii. More general: improving governance, rationality, moral enhancement, space colonisation (and research on mitigation options…)
    d. Question: how cost-effective (expected good per dollar) are our best mitigation options? In particular, how do those options compare to the value of the best non-x-risk "EA interventions"?
    e. Basic idea: premature extinction would be *so extremely bad* that reducing its risk even by a *tiny amount* is highly valuable (and potentially more valuable than e.g. the best immediate life-saving/health-improvement interventions).
2. Warm-up exercise: My thatched roof
    a. I have a thatched roof. Thatches sometimes catch fire. My insurers recommend that I buy a "thatch monitoring system", but these systems are quite expensive. Is it worth the cost?
    b. Naïve decision theory (maximin): do whatever makes the worst-possible outcome as good as possible.
        i. This is too risk-averse; the cost must be more relevant than that.
        ii. Anyway, this theory would probably tell me to ignore the possibility of thatch fire (my house burning down isn't the *worst possible* outcome).
    c. Expected-value approach: all things are quantitative…
        i. I have to decide

1. *How much* worse it would be if the thatch caught fire/house burnt down than if it didn't. (ΔV)
2. *By how much* the monitoring system would decrease the probability of fire. (Δp)
3. How much I value the money that I'd have to spend to buy the system. (c)

ii. Expected value theory tells me to buy the system iff the "expected" (i.e. probability-weighted) value of buying exceeds that of not-buying.
1. I.e. iff the expected value difference EV(buying) – EV(not buying) is positive.

iii. The difference in expected value is: $\Delta p.\Delta V - c$.

iv. This could go either way, but is more likely to be positive (to favour taking the precaution) as ΔV becomes larger.

3. A simple model of the x-risk mitigation decision
   a. Divide logical space into X ("premature extinction" of sentient life, defined as e.g. extinction within 100 years) and not-X (continued survival).
   b. The probability of X – call this probability p – depends on our actions.
   c. There are two available actions:
      i. Not-M (don't mitigate): The status quo. Then there is a probability p of X.
      ii. M (mitigate x-risk): Relative to the not-M case, we reduce p by a fixed amount – call this amount Δp – for a certain cost c.
      iii. Assume that the choice between M and not-M does not affect the probability of any event *conditional* either on X or on not X (i.e., does not affect either of the functions p(.|X), p(.|not-X)).
   d. Thus there are four possibilities, with [expected] values as follows:

	X	Not-X
Not-M	V_X	$V_{not\text{-}X}$
M	V_X - c	$V_{not\text{-}X}$ – c

   e. According to the expected-value approach to action under uncertainty (recall, from week 3), the subjectively best/right option is the one with the highest expected value.
      i. The expected value calculations are:

$$E[V(\text{not-M})] = p\,V_X + (1\text{-}p)\,V_{not\text{-}X}$$

$$E[V(M)] = (p - \Delta p)\,V_X + (1\text{-}p+\Delta p)\,V_{not\text{-}X} - c$$

      ii. The expected value of mitigating exceeds that of not mitigating by the [possibly negative] amount

$$\Delta p\,(V_{not\text{-}X} - V_X) - c.$$

4. Comparative question (crucial for 'EAs'): assuming (for the sake of argument) that the value of mitigation is higher than that of "doing nothing", is it also higher than the value of the best non-x-risk intervention?

a. Beckstead's (2012) benchmark: With a $20bn budget one could 'save' 12.5 million (already existing) lives.

5. Evaluating the expected benefit of x-risk mitigation
    a. The value of Δp (for a fixed c) is an empirical question.
        i. Beckstead's conservative lower-bound estimate (from studies of asteroid deflection possibilities): $\Delta p = 5 \times 10^{-7}$, for a cost c = $20bn.
    b. The value of $(V_{not\text{-}X} - V_X)$ depends complicatedly on both empirical and evaluative issues – in particular, evaluative issues in population axiology…
6. …Simplest case: Totalism (and temporally additive theory of lifetime well-being)
    a. Totalism:
        i. Each person i (existing in state of affairs x) has a certain *lifetime well-being score* $w_i(x)$, measuring how well person i's life goes in x.
        ii. The value of a state of affairs x is the sum of the well-being levels of all persons who ever live in x: $V(x)=\Sigma_i\, w_i(x)$.
    b. Notable feature of totalism: facilitating the existence of "enormous" numbers of additional lives tends to be "enormously" important (provided the average well-being level of those lives is positive).
        i. In particular: creating extra well-being by creating extra happy lives has the same value as creating extra well-being by improving (e.g. extending) the lives of existing people. (For fixed well-being quantities.)
    c. Example calculation:
        i. Assume (for simplicity) a simple additive theory of lifetime well-being: A person's lifetime well-being level is the sum (or time-integral) of another quantity, "momentary well-being", across the time-periods in her life.
        ii. If V_X corresponds to sentient life ending in 2100 while $V_{not\text{-}X}$ involves an extra 1 billion years of sentient life, with the same average well-being levels as existing lives, then $V_{not\text{-}X} - V_X$ is *ten million times* the value of all the well-being contained in the present century.
        iii. Suppose that some mitigation action M could reduce the probability of premature extinction by one in a million. ($\Delta p = 10^{-6}$.)
            1. Then Δp $(V_{not\text{-}X} - V_X)$ is ten times the value of *all the well-being contained in the present century*.
            2. Likely to vastly outweigh the cost c, for any remotely feasible x-risk mitigation project. (Relevant to the 'absolute' question.)
            3. Also likely to dwarf the benefits of non-x-risk projects that can be carried out for the same cost. (Saving 12.5 million lives?)
7. Interlude: Population axiology (see Greaves, 'Population axiology', for an overview)
    a. The question of population axiology: Assign values to states of affairs, when some pairs of states of affairs in the domain differ from one another over how many people ever exist. (Pulls apart e.g. average and total 'utilitarianism'.)
        i. (Closely related: the bigger question of population *ethics*.)
    b. [If the lifetime well-being scale has the structure of the real numbers, then] Totalism notoriously entails the "*Repugnant Conclusion*": For any state of affairs A [no matter how good], and any 'barely worth living' lifetime well-being level $\varepsilon>0$, there is a better state of affairs in which no-one has a well-being level higher than ε.

c. On the other hand, Averagism entails e.g. the "Sadistic Conclusion": It can be better to add people with negative well-being than to add [different numbers of] people with positive well-being, starting from a common baseline scenario.
d. "Impossibility theorems" of population axiology/population ethics: For various collections of intuitively compelling conditions ('avoid the Repugnant Conclusion', 'avoid the Sadistic Conclusion', etc.), it is provably the case that no population axiology satisfies all the conditions simultaneously. (So some intuition has to give.)
    i. Could be (and has been) used to defend Totalism. But there are other options.
e. One salient family of alternatives: "Person-affecting theories"
    i. Basic person-affecting intuitions:
        1. The "person-affecting principle": For all states of affairs x, y: x is not better than y unless it is better *for some person*.
        2. The "denial of existence comparativism"/"negative answer to the existential question": x cannot be better than y for S unless S exists in both x and y. (You don't (comparatively) benefit a person by creating her.)
    ii. Consequence for the analysis of extinction risk: The vast number of additional possible future lives is irrelevant for the evaluation of $V_{not\text{-}x} - V_x$. This quantity is just a matter of
        1. the amount by which continued-existence is better for already-/independently- existing people, and
        2. non-welfarist considerations (the intrinsic value of the continued existence of civilisation, art, scientific achievements…?).
    iii. It is notoriously difficult to actually *formulate* a remotely plausible person-affecting theory. Some initial attempts:
        1. Presentism: Goodness = total well-being of presently existing people.
            a. Totally implausible if there are future people *who are going to exist regardless of what one does*.
        2. Actualism: Goodness = total well-being of actually existing people (including people who don't yet exist, but who will *in fact* exist in the future).
            a. Violates 'axiological invariance'
        3. Necessitarianism: Goodness = total well-being of people who exist (at any time) regardless of which choice is made in the present decision.
            a. Choice-set dependent
    iv. None of these looks very good. But perhaps we could/should keep trying…

8. Objection: The expected value of the future is negative.
    a. This is so on (a) some pessimistic empirical hypotheses, (b) some depressing theories of well-being. (E.g. "Frustrationism": count only frustrated desires, and count these negatively.)
    b. The conclusion from this assumption is likely to be that it is overwhelmingly important to *increase* x-risk (for precisely analogous reasons).

        i. Not a way of resisting the more general conclusion that the most cost-effective interventions are x-risk ones!
9. Sensitivity analysis: What if we aren't Totalists?
    a. Beckstead's insensitivity thesis: quite a lot of the conclusion (that mitigating x-risk is enormously valuable) remains, on all but the most implausible alternative approaches to population axiology.
        i. To emphasise this point, Beckstead's exposition does not actually proceed by assuming Totalism in the first place. (Instead: "Period independence", "Additivity", "Temporal impartiality".)
        ii. 'All but the most implausible':
            1. A "strict" person-affecting view that assigns *zero intrinsic value* to the creation of additional future lives (and thus to the longer survival of sentient life) is implausible. The more plausible versions assign *some* intrinsic value to this, but less than they do to the improvement of additional lives. But then, for plausible numbers, Beckstead's basic conclusion is likely to remain.
            2. 'Variable value' theories violate an 'independence' condition (cf. "the Egyptology objection" to Averagism.)
            3. "Critical level" theory leads to substantially the same conclusions.
    b. Resisting Beckstead's insensitivity thesis
        i. The most plausible person-affecting view is probably one that *hasn't been stated yet*, not the 'softened' version of presentism/actualism/necessitarianism that Beckstead deals with. And might well assign a value to continued survival that is (i) significant, but (ii) non-enormous even for enormous potential futures. (?)
10. Sensitivity analysis (II): What if we don't accept expected value theory?
    a. Note that commonly suggested ways of departing from the expected-value approach are in some sense *more risk averse* than expected-value theory.
        i. 'More risk averse' *with respect to* the value scale we have previously assumed: maximise the expected value of f(V) rather than that of V, where f is an increasing but concave function.
            1. (Still consistent with "the axioms of expected utility theory".)
        ii. A different sense of 'risk averse':[34] our approach to uncertainty does not have the structure of maximising the *expectation value* of *any* quantity. Instead, for cases of two possible outcomes use e.g. V(worst possible outcome) + r(p(better outcome))[V(better outcome) – V(worse outcome)], where r is increasing but convex.
            1. (Violates EU theory.)
    b. On the other hand, maybe some departure in the other direction is also warranted:
        i. "Pascal's mugging": A mugger approaches you. He has no weapon, but exhorts you to hand over your wallet: "In return, I will give you any finite

[34] This is the approach of "risk-weighted utility theory." See e.g. Buchak, *Risk and Rationality* (OUP 2013). For a paper-length summary see "Risk and tradeoffs", https://philosophy.berkeley.edu/file/754/Buchak_Risk_and_Tradeoffs.pdf

amount of utility that you ask for. I'm able to do this because I have secret powers. Now, you might think it's extremely unlikely that I'm telling the truth here, but surely you have *nonzero* credence that I am; and if so, you only have to stipulate a sufficiently high utility reward, and then handing over your wallet will have positive expected utility for you."

ii. Possible responses
    1. Reject EU theory for very small-probability, high-stakes scenarios, and err in the direction of neglecting extremely unlikely possibilities.
        a. This would undermine the argument for prioritising x-risk.
    2. Bite the bullet (i.e. agree with the mugger)
    3. Bounded utilities
    4. Zero credence
    5. Sufficiently fast-diminishing credence that the mugger is able to supply increasingly large utilities

iii. More general (but nebulous) worry: Taking the argument to its logical conclusion would lead one to base just about *all of one's decisions* on considerations of effects on x-risk (at least insofar as one is altruistic - e.g. views on governance and public policy.) This seems "fanatical".
    1. How much of an objection is this?

11. Objection: What if non-x-risk interventions have knock-on benefits that continue indefinitely? Couldn't they then turn out to be better than x-risk ones?
    a. Making this thought more precise: On the totalist calculation,
        i. The expected value of reducing x-risk by Δp was given by Δp.E(L).w, where w = average well-being in a future life.
        ii. Suppose e.g. that some non-x-risk intervention would, besides its immediate and direct benefit, make future lives better by an average amount Δw per future life (where Δw is approximately independent of the number of future lives). Then the non-x-risk intervention would have an additional expected value of E(L).Δw.
        iii. Both of these quantities scale linearly with E(L). The comparative question now is just whether Δp.w or Δw happens to be higher (could go either way).
    b. Beckstead's reply: This model doesn't apply to knock-on effects that merely *speed up progress* by a fixed amount of time, since there are limits to growth. And that's the most plausible scenario for e.g. life-saving interventions.
    c. But there could be other ways of speeding up progress, and other types of knock-on benefit. Mightn't those turn out enormously valuable?
        i. Those who favour prioritising "existential risk" actually agree with this…
12. …Beyond "extinction"
    a. Theorists seek to define "existential risk" in a way that *includes* (at least some) other cases of possible ongoing impacts leading to very large drops in expected value, despite possibly small probabilities.[35]
    b. Some suggested definitions

[35] For discussion, see Ord and Cotton-Barratt, "Existential risk and existential hope", available from http://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf

i. Bostrom: “An existential risk is one that threatens the premature extinction of Earth-originating intelligent life *or the permanent and drastic destruction of its potential for desirable future development*.” (Bostrom, cited in Ord & Cotton-Barratt; emphasis added)
ii. Bostrom and Circovic: “An existential risk is one that threatens to cause the extinction of Earth-originating intelligent *life or to reduce its quality of life (compared to what would otherwise have been possible) permanently and drastically*.” (*Global catastrophic risks*, p.4; emphasis added)
iii. Ord and Cotton-Barratt: “An existential catastrophe is an event which causes the loss of a large fraction of expected value.”
iv. In Beckstead’s PhD thesis, the central claim is that “what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.”

c. On some of these (but not others), risks of a condition that would cause mild pain to every future creature throughout its existence would count.
d. Open question: what exactly is the most fruitful definition for “x-risk theorising”?

13. Isn’t *increasing existential hope* just as important as *decreasing existential risk,* by the lights of this sort of argument?
    a. E.g. space colonisation
    b. Answer (by the lights of the x-risk arguments themselves):
        i. In principle, yes
        ii. “Status quo bias” is a mistake[36]

[36] Bostrom and Ord, “Status quo bias in bioethics”

Foundations of Effective Altruism
Frank Arntzenius and Hilary Greaves
Hilary Term 2017
Week 8: EA evaluations and cluelessness

1. Randomised controlled trials (RCTs)
    a. "Evidence-based medicine": Don't just base treatment or prioritisation decisions on 'hunches' about what seems the best thing to do, or on 'experience'. Do proper experiments to quantify effects.
        i. 'Proper' experiments: ideally, 'randomised controlled trials'.
    b. Naïve observational studies: observe the degree to which the desired effect E tends to be *correlated* with the proposed intervention I.
        i. Examples:
            1. Connection between smoking and cancer
            2. Effectiveness of surgery for
                a. Cirrhosis of the liver (with treatment decisions made by doctor and/or patient)
                b. Coronary bypass surgery (using historical controls)
        ii. The problem of confounding factors:
            1. Mere correlations between E and I could be due to
                a. a causal link from I to E (the desired case)
                b. a causal link from E to I (not applicable in our cases of interest), or
                c. a common-cause structure (a very real possibility).
            2. Observational studies cannot distinguish between these possibilities.
            3. Only the first would justify prescribing/prioritising I on grounds of E.
    c. Randomised controlled trials: Find a large population to experiment on. *Randomly* assign some of the participants to 'treatment' vs 'control' groups. Measure the extent to which the incidence of E is higher in the 'treatment' than in the 'control' group.
        i. The randomness eliminates the possibility of confounders…
        ii. …*except* those that are included in the treatment procedure. (E.g. the placebo effect.)
            1. To fix this: Replicate as many aspects of the treatment procedure as possible, except the intervention I itself, in the 'control' group. (Placebo-controlled trials; double-blinding.)
    d. In the medical literature, the use of RCTs has overturned many observational-study 'results' that previously seemed quite plausible.
    e. Problems with RCTs
        i. Availability (e.g. there are few trials for safety of drugs on pregnant women).
        ii. The problem of external validity: to what extent do the results generalise beyond the context in which the trial was carried out?
2. Application to EA

a. An unsuccessful intervention: Playpumps[37]
    i. Many people conclude from this that aid doesn't work *in general*.[38]
b. More optimistic conclusion: we just have to be careful (here as everywhere else).
    i. In particular: *RCTs* can be applied to questions outside of medicine too, to test the effects of proposed philanthropic interventions. (Characteristic of 'J-PAL': https://www.povertyactionlab.org/ .)
c. Evaluating deworming, e.g. SCI
    i. E.g. Miguel and Kremer's study of deworming[39]
        1. Question: "To what extent does deworming children boost school attendance?"
        2. Approach: Select 75 primary schools in rural Kenya. Allocate schools to 'treatment' (mass deworming administered at school) and 'control' (no treatment), by a modified pseudorandom process (designed to maximise across-group homogeneity while retaining effective randomness). Collect subsequent school attendance records in schools from both groups.
        3. Result: Mass deworming boosts school attendance by one child-year per $3.50 deworming cost.
d. Evaluating distribution of long-lasting insecticide-treated nets (LLINs), e.g. AMF[40]
    i. Question: "How many deaths are averted per 1000 children supplied with a LLIN (for children under age 5)?"
    ii. Answer suggested by RCTs: 5.5

3. The importance of indirect effects
    a. What EAs ultimately seek to maximise is *total* good done per dollar donated.
    b. Effects measured via RCTs are necessarily only a small part of an intervention's total effects.
    c. Things not counted:
        i. Further effects of the measured "effect". E.g., if the RCT stops at 'additional years of schooling' or 'lives saved': knock-on effects on e.g. employment prospects, economic growth, biodiversity, fertility rate and long-run population…
        ii. Any 'side-effects' of the effect being focussed on. E.g. political effects of carrying out the intervention.[41]
    d. It's plausible (all but inevitable?) that the uncounted effects constitute the majority of total good (or harm) done.
    e. The importance of this

[37] See e.g. the opening pages of MacAskill, *Doing good better.*
[38] See e.g. Moyo, *Dead aid: Why aid isn't working and how there is another way for Africa*; Easterly, *The white man's burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*
[39] Miguel and Kremer, "Worms: Identifying impacts on education and health in the presence of treatment externalities", available from http://cega.berkeley.edu/assets/cega_research_projects/1/Identifying-Impacts-on-Education-and-Health-in-the-Presence-of-Treatment-Externalities.pdf.
[40] http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000363.pub2/full
[41] See e.g. Clough, "Effective altruism's political blind spot", *Boston Review*

        i. For *cost-effectiveness* analysis *within a single cause*, the problem of *knock-on* effects can be set aside. (These only affects the value one should attach to the measured effect.)
        ii. But for the purposes of (i) cost-benefit analysis and (ii) cross-cause cost-effectiveness analysis, knock-on effects are important.
        iii. And side-effects are important for all types of analysis.
    f. Applies most to global poverty/health interventions; less to animal suffering and x-risk.
4. Systemic change
    a. There are various policy reforms that would be enormously valuable (in expectation) *if* we could achieve them.
        i. E.g. Clean trade, pharmaceutical prices, political oppression, corruption, open borders…
    b. It is usually impossible to do RCTs for systemic-change interventions, because there is no relevant 'large population' to experiment on (even if such experiments would be ethically acceptable).
    c. As a result, any estimates of the expected efficacy of systemic-change interventions (and, to a lesser extent, of the value of successful reforms) are necessarily highly subjective.
    d. Some EA organisations recommend/fund systemic-change interventions nonetheless (e.g. the Open Philanthropy Project). Others tend to recommend only interventions for which the supporting evidence is more robust (e.g. GiveWell).
5. Cluelessness (a brief summary)
    a. Basic question of this paper: To what extent (and in what ways) does lack of information about the long-term consequences of our actions paralyse rational decision-making, given a concern with total (as opposed to immediate/direct/foreseeable/etc.) good done?
    b. Lenman's answer: Totally, given consequentialism.
    c. Narrative in my paper:
        i. (This isn't really about consequentialism.)
        ii. There's no problem in the simple cases that Lenman and others have worried about, because those are cases in which (for rational credences, or anyway for my credences) the unforeseeable effects make zero contribution to *expected* value.
            1. Although Lenman is right about the 'objective' case.
        iii. But the EA case is relevantly different, as here we face 'complex' rather than 'simple' cluelessness: "some reasons pointing in each direction, and unclarity about how to weigh these reasons up against one another."
            1. Relatedly, many people seem(?) to suffer from some sort of decision aversion/paralysis in these cases.
        iv. It's unclear how to analyse 'complex' cluelessness.
            1. Lack of guidance from theory re which credence function to adopt?
            2. Imprecise credences? (If so, which decision theory for imprecise credences?)
            3. Non-resilient credences? [Not considered in my paper.]

v. Relatedly: It's unclear whether there is any sense in which lack of solid information about further effects (in 'complex cluelessness' cases) can 'paralyse' rational decision-making.

6. A bit more on non-resilient credences
    a. Two cases of 50% credence
        i. Case 1: The urn contains 50 red and 50 black balls. A ball has been drawn at random. What is your credence that it was red?
        ii. Case 2: The urn contains either 99 red and one black ball, or vice versa. You have equal credences in these two possibilities. A ball has been drawn at random. What is your credence that it was red?
        iii. Your 50% credence in 'red' is more resilient to possible additional evidence in the first case than in the second.
    b. Non-resilient credences could comprehensively lead to a form of decision aversion: one would rather acquire new evidence than make a forced choice now, if the stakes are at all high, and provided that the new evidence is not itself too costly.
    c. This might be part (?) of what's going on in the EA cases, as a *psychological* matter. (?)
    d. But note that it is in general *irrational* to prefer a fourth option D ('do nothing') over both A and B, on the ground that some *other* option C ('acquire more information before deciding between A and B') is better than both A and B.