Effective altruism, dynamic effects, and coordination

Stefan Schubert, Centre for Effective Altruism

Effective altruism

- Effective altruism: idea and movement whose goal is to *do the most good*

- Precisification (MacAskill): promoting welfare using *evidence and reason*

- Compatible with a range of ethical theories and methods for doing good

Cause neutrality and prioritisation

- Distinctive features: *cause neutrality* and *means neutrality*

- Entail *prioritisation*

- Cf Lionel Robbins: economics ”studies human behavior as a relationship between limited resources and unlimited wants which have alternative uses”

- Economic thinking hitherto strongly under-utilitized among altruists.

Effective altruist work

- In practice, effective altruists work on:

- Global poverty, animal welfare and existential risk mitigation

- Movement building and improvement (meta-work)

- Methods include charity rating (GiveWell) and career advice (80’000 Hours)

Effective altruism as evidence-based philanthropy

- Effective altruism viewed as introducing evidence into philanthropy. Steven Pinker:

“In many spheres of life, observers are replacing gut feelings with quantitative analysis. Sports have been revolutionized by *Moneyball*, policy by *Nudge*, punditry by *538.com,* forecasting by tournaments and prediction markets, philanthropy by effective altruism, the healing arts by evidence-based medicine.”

- Question: what kind of evidence?

Effective altruism and evidence

- In principle, EAs are open to many different uses of evidence

- In practice, effective altruism has been associated with (e.g. Angus Deaton, Daron Acemoglu, Iason Gabriel):

a) Randomized controlled trials

b) Focus on direct, clearly observable effects of actions

Effective altruism and evidence

- True that within global poverty and health, many EA recommendations are RCT-backed

- Against Malaria Foundation, Give Directly

- Outside global poverty, this is much less true (James Snowden):

- E.g., existential risk, career advice, movement building

Effective Altruism and unsalient effects

- Other trend in EA: strong focus on *indirect effects* and *effects of others' action*

- What are the far-future effects of global poverty interventions?

- To what extent will what looks like high-impact problems today become less impactful thanks to future work on them? (Future non-neglectedness)

- To what extent will an EA doctor replace another doctor, who would have done an equally good job? (Replaceability argument)

- Neither EAs themselves nor their critics tend to highlight this feature in discussions about effective altruism.

Effective Altruism and unsalient effects

- Will here focus on effects of others' actions
- Serve to give us a better picture of EA's relation to evidence
- Important if effective altruism is in some sense the philanthropy counterpart of evidence-based medicine
- Also important in its own right

Effects of others’ actions

- Effective altruists are better than other altruists at taking these effects into account.
- Focus on *counterfactual impact* – as opposed to direct impact
- ”What changed, because of my actions?” rather than
- ”What direct impact did I have?”
- Could plausibly could do better, as we shall see.

Effective altruism and effects of other’s actions

Questions ordered after the degree to which effective altruists have discussed them and integrated them into their recommendations.

1) Neglectedness
2) Replaceability argument
3) Coordination problems between effective altruists
4) Allocating work across time/future non-neglectedness
5) Dynamic effects on controversial issues

Theory and practice of Effective altruism

- These problems arise from effective altruist *practice*.
- We need to decide what to do, and what to recommend others to do, in order to do the most good.
- Will give some tentative thoughts on the problems, but no full solutions.
- We plan to set up an academic institute for effective altruism at University of Oxford, which will study them in more detail.

Neglectedness

- Effective altruists find how *neglected* a problem is highly important

- If the marginal returns on working on a problem are diminishing, neglectedness is a *reason to* work on a problem

- If the marginal returns on working on a problem are increasing, neglectedness is a *reason against* working on a problem (cultured meat?)

- Major reason why more effective altruists work on AI safety thanclimate change

The replaceability argument

- Naively, it might be thought that entering medical school might be a good option for an altruist
- However, if you decide not to go into medical school, you will be *replaced* by someone else
- Unless that person is less skilled than yourself, your counterfactual altruistic impact can become zero
- Cause: quantity restriction

The replaceability argument

- The replaceability argument can in general be analyzed using a supply and demand framework (Rossa O’Keefee-O’Donovan)

- Supply and demand elasticities influence levels of replaceability

- The more elastic labour supply is, the greater the replaceability effects are

The replaceability argument

- E.g. since labour supply of charity workers probably is inelastic, replaceability effects probably are weak

- Earning to give: ”earn all you can, save all you can, give all you can”

- If an EA decide not to become, e.g. a quantitative trader, they will most likely be replaced by a non-altruist.

- For this and other reasons earning to give may beat entering medical school and other traditional altruistic endeavours.

The replaceability argument

- Possibly perverse effect (Michael Dickens): case to be made that:
- Non-altruists should be persuaded to do altruistic work
- Altruists should do non-altruistic work and use their earnings altruistically.

The replaceability argument

- Though perhaps obvious, it is often ignored by other altruists
- EA insight mainly due to their conceptual sophistication
- Knowledge of philosophy and economics
- EA have somewhat toned down the importance of replaceability. More research is needed.

Coordination and replacement regarding EA donations

- Suppose that you are donating to a particular charity A

- That leads to another donor changing their donations from A to another charity B

- Your counterfactual impact is thereby on B rather than A

- B may be less effective than A (Iason Gabriel). Decreases your counterfactual impact

- Additional donation supply may lead to the creation of new charities. Increases your counterfactual impact.

Coordination problems between EAs regarding donations

- You may have strong ethical reasons to prefer A over B.

- Suppose that there is another charity C which the other donor strongly disprefers to A and B

- You would prefer donating to A over C if it were not for the donation replacement issue.

- In this case, you may donate to C even though that reduces total moral preference satisfaction

Coordination problems between EAs regarding donations

- Moral trade (Toby Ord, 2015) can in principle solve this issue

- You can come to an agreement with the other donor that you both donate to A, which both of you like, rather than to B and C

- Problem of *counterfactual trust*: how do I know what other donors would have done if were not for the possibility of me donating?

Coordination problems between EAs

- We have only skimmed the surface of these questions
- Much more research needed on coordination, both regarding donations and direct altruistic work
- Ord’s initial work on moral trade suggests that coordinating altruistic efforts effectively may pose unique challenges relative to coordinating self-interested efforts.

Allocating work across time

- Suppose that we want to solve a problem that is currently highly neglected – e.g. AI risk or biotechnological risks.
- This could cause an existential catastrophe at some unspecified time *t* in the future.
- Risk-mitigating work may in principle be carried out at any time before *t*.
- (Some kinds of work closer to *t* may be more effective - near-sightedness)

Future rationality

- Deep uncertainty over how much work others will carry out on the problem in the future.

- *Attribution of rationality to future actors* may imply that they will do considerably more work than current trends suggest

- May occur due to future event (Andrew Snyder-Beattie)

- ”Warning shots” (minor catastrophe) or ”warning signs” (near-misses) may increase risk-mitigation work (e.g. regarding bio-risks)

Future rationality

- Warning shots/signs idea assumes relatively low levels of rationality

- Theoretical argumentation may be enough to trigger an appropriate response (has to some extent been happened regarding AI safety)

- Leads to lower levels of *future neglectedness*

- Suggests that current work on such problems is less impactful than it would be were it not for this factor

The future rationality consideration

- Easy to overlook this factor

- In general, there may be less variance in terms of impact between different problems than a naive EA analysis suggests

- As an ”efficient market for altruistic interventions” develop, the ”low-hanging fruit” are increasingly picked.

- May increase the relative value of building general capacity relative to solving specific problems

Overuse of the future rationality consideration

- People might delay risk-mitigation work, thinking that future actors will solve the problem

- ”Bystander effect” across time

- Especially dangerous if the costs of acting and omitting to act are asymmetric

Ersatz version of the future rationality consideration

- Many predicted problems actually get solved
- People observe this, and conclude that warnings often are exaggerated
- However, problems are often solved precisely*because* people act rationally to solve the problem.
- ”Self-defeating predictions”
- People either fail to posit a mechanism, or posit another mechanism

Ersatz version of the future rationality consideration

- Dangerous if it leads to thinking that problems will be solved automatically

- Partly due to ambiguity regarding predictions/warnings: are they to be interpreted as saying:

- ”Problem X will cause Y harm at point *t*” or

- ”Problem X will cause Y harm at point *t unless we take action*”

Effective altruism and the future rationality consideration

- Rapid unpredicted expansion of AI safety-work suggests EAs have overlooked this factor
- If EA continue to expand rapidly, this factor will be very important.
- More research needed
- We need to become better at incorporating existing ideas into current prioritisation decisions

Dynamic effects on controversial questions

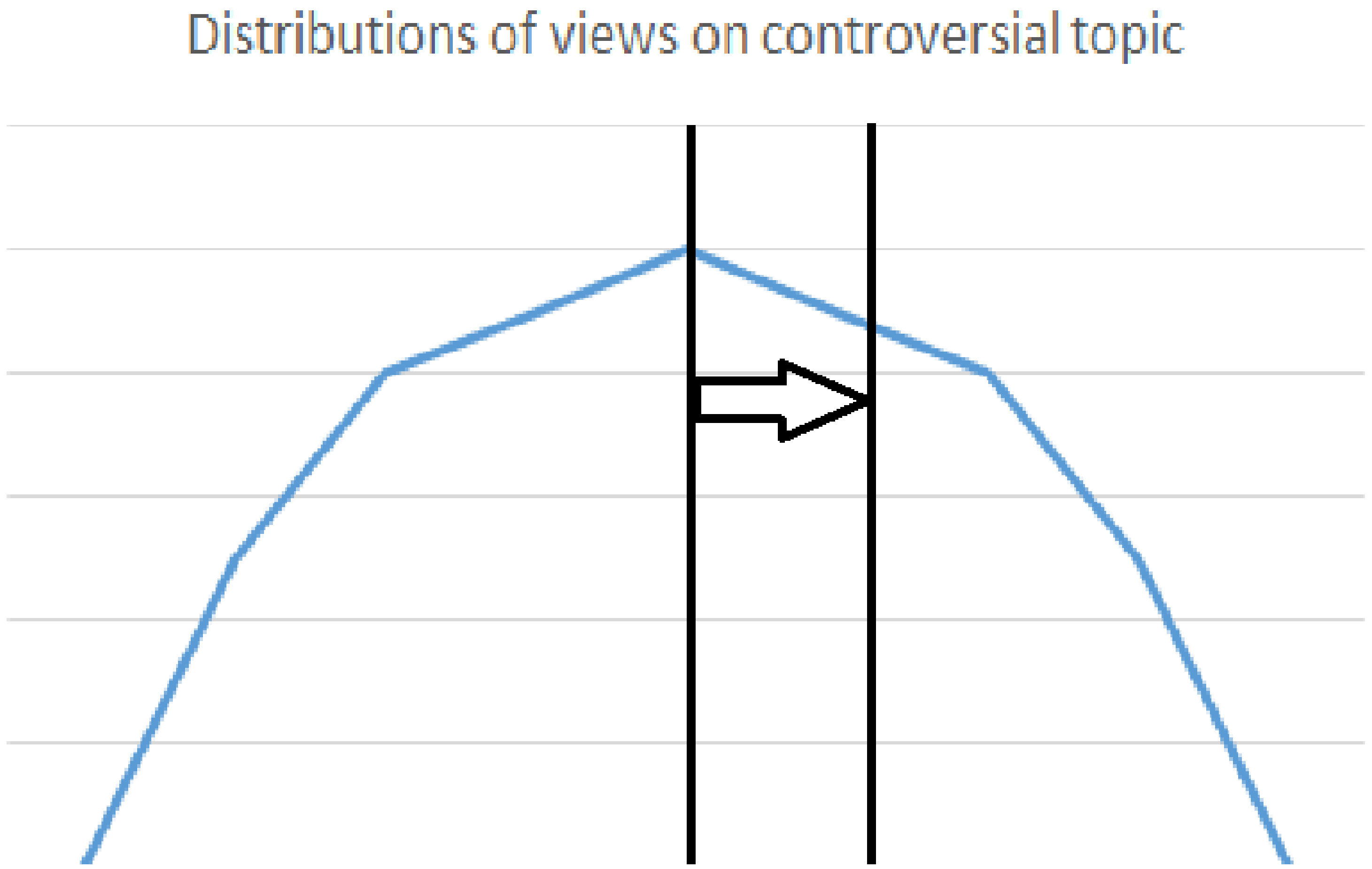
- Questions like, e.g. immigration, development aid, criminal justice reform, LGBT laws

- Suppose that current policies on these issues are close to the views of the majority of the population

- Suppose that you succeed to push them in a progressive direction, thanks to, e.g. lobbying

- What will the effects be?

Dynamic effects on controversial questions



Dynamic effects on controversial questions

Highly static (naive?) scenario: no reaction from voters

Moderately dynamic scenario: voters react on existing views, and push back against the change – backlash

(Note that you may still capture some value before that happens)

Highly dynamic scenario: voters' change their views – either in line with your views, or against your views

Dynamic effects on controversial questions

- Functionalism – society is in equilibrium, and many attempts to change it will suffer from these effects (Brexit - immigration?)

- Lenin: ”the worse the better”:

The further the ruling powers push policies away from the views of the majority, the greater the chance of revolution (population view change).

- Elite avantgardism: policy change, e.g. on LGBT issues, will lead to view change

Dynamic effects on controversial questions

- Effective altruists have not taken this factor into account to the extent that we should
- May push away from lobbying and towards working on changing voters' views
- More research is needed

Effective altruism, evidence, and practice

- Many of these problems are exceedingly hard
- Hard to obtain rigorous evidence on them
- Possible attitude: we should only take effects on which we have rigorous evidence into account
- Increasingly common EA view: focus on maximising expected utility dictates that effects on which we have weak evidence should be taken into account.

Effective altruism, evidence, and practice

- We have no choice but to speculate about highly uncertain effects of others’ actions, indirect effects, and far-future effects

- Salient feature of, e.g. Future of Humanity Institute (EA-org, University of Oxford)

- Standard academic attitude is to shun away from such speculation.

- Effective altruists’ different attitude stems from our focus on practical action

Conclusions

- Effective altruism does not focus on randomized controlled trials and direct effects in the way critics suggest

- Rather, we are more interested in speculative interventions, and highly indirect effects, than most

- For instance, we are trying hard to reason about the effects of others' actions, and how they should affect prioritisation

References

- Daron Acemoglu, Angus Deaton, Peter Singer and others: ”The Logic of Effective Altruism”.

- Michael Dickens: ”Altruistic Organizations Should Consider Counterfactuals When Hiring” EA Forum

- Iason Gabriel: ”Effective Altruism and Its Critics”.*Journal of Applied Philosophy*

- John Halstead et al. ”Effective Altruism: An Elucidation and A Defence”. Unpublished.

- Will MacAskill: ”Replaceability, Career Choice, and Making a Difference” *Ethical Theory and Moral Practice*

References

- Rossa O’Keefee O’Donovan: ”What does economics tell us about replaceability?” 80’000 Hours

- Toby Ord: ”Moral Trade”.*Ethics*

- Stefan Schubert: ”Sleepwalk bias, self-defeating predictions and existential risk”. LessWrong

- James Snowden: ”Why effective altruism used to be like evidence-based medicine. But isn’t anymore” EA Forum

- Andrew Snyder-Beattie: ”Implications of ’Warning Shots’ on Existential Risk Reduction” Unpublished.